AD________________

Award Number: DAMD17-96-1-6288

TITLE: Reaching Rural Mammographers for Quality Improvement

PRINCIPAL INVESTIGATOR: Nicole Urban, ScD

CONTRACTING ORGANIZATION: Fred Hutchinson Cancer Research Center
Seattle, Washington 98104-2092

REPORT DATE: October 1999

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release; distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	**2. REPORT DATE** October 1999	**3. REPORT TYPE AND DATES COVERED** Annual (16 Sep 98 -15 Sep 99)

4. TITLE AND SUBTITLE
Reaching Rural Mammographers for Quality Improvement

5. FUNDING NUMBERS
DAMD17-96-1-6288

6. AUTHOR(S)
Nicole Urban, ScD

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Fred Hutchinson Cancer Research Center
Seattle, Washington 98104-2092

E-MAIL:
urban@fhcrc.org

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release;
distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

The Fred Hutchinson Cancer Research Center, the University of Washington, and the Washington State Department of Health are collaborating to develop and implement a mammography quality improvement program (MQIP), in order to demonstrate its feasibility and effectiveness for dissemination. The MQIP emphasizes continuous quality improvement (CQI) in film interpretation, within the context of a comprehensive program designed to meet the requirements of the Mammography Quality Standards Act (MQSA) of 1994.

During the third year of funding, project investigators and staff focused on finalizing and conducting the CQI, recruiting facilities to the Mammography Tumor Registry, cleaning and linking data and providing audit reports to participating facilities. Additionally, evaluation of technologist training took place as part of the certification function

A no-cost extension is being requested to complete the project and conduct overall evaluation.

14. SUBJECT TERMS
Breast Cancer; mammography; quality assurance

15. NUMBER OF PAGES
74

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

___ Where copyrighted material is quoted, permission has been obtained to use such material.

___ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

___ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

NA In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

X For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

N/A In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

N/A In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

N/A In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Nicole Urban 8/23/99

PI - Signature Date

Annual Report for Grant DAMD17-96-1-6288

September 16, 1998 - September 15, 1999
Year 03

Reaching Rural Mammographers for Quality Improvement

Nicole Urban, ScD
Principal Investigator

Table of Contents

INTRODUCTION

The Fred Hutchinson Cancer Research Center (FHCRC), the Department of Radiology at the University of Washington (UW), and the Washington State Cancer Registry (WSCR) at the Department of Health (DOH) are working together to develop a comprehensive Mammography Quality Improvement Program (MQIP) designed for easy dissemination, particularly in rural areas. The MQIP emphasizes improvement in film interpretation, within the context of a comprehensive program designed to meet the requirements of the Mammography Quality Standards Act (MQSA) of 1994. The project focuses on rural areas because these communities have been identified as being underserved by public health research.[1, 2, 3] Additionally, there may be cause for concern about the quality of care offered in rural areas.[4]

The MQIP is a demonstration project that consists of four basic functions. The *surveillance* function employs routine systematic monitoring of measurable outcomes of screening mammography, including sensitivity, specificity, and positive predictive value. As part of the *audit* function, the project also identifies for mammographers their false positive and false negative cases, so that they can improve quality through review of their own films. In addition, the project provides continuing education for radiologists, and training for technologists, as required by MQSA as well as training for registrars. This is its *certification* function. The *continuous quality improvement* function provides immediate feedback following a radiologist's interpretation of practice films selected for their educational value. The MQIP is comprehensive, and will ensure that participating facilities are in compliance with evolving accreditation rules.

The MQIP builds on another project funded through the National Cancer Institute that is being conducted at the FHCRC entitled the Washington Mammography Tumor Registry (MTR) (Nicole Urban, P.I.). The MTR is a registry of mammography data obtained from facilities in Washington State, which is linked to tumor data obtained from the WSCR and the Puget Sound Cancer Surveillance System. The purpose of this registry is to provide a resource for research into mammography performance and breast cancer in addition to offering informational reports to participating radiologists and facilities. The MTR will be used to accomplish the surveillance and audit functions of the MQIP.

A research study is being conducted within the MQIP demonstration project. The **primary research objective** is to determine if the CQI program can increase the accuracy with which mammographers interpret films. **Secondary research objectives** are to 1) determine inter-rater variability in film interpretation in a set of films selected for their teaching value, before and after implementation of the CQI program; 2) determine post-CQI intra-rater variability in film interpretation; 3) determine if digitized films can be interpreted with the same accuracy as can high-quality copies of films; and 4) determine if the accuracy with which films are interpreted depends on covariates, the age of the woman being of particular interest. The availability of comparison films will also be considered as a covariate.

This three-year project is currently at the end of the final year.

BODY

Eighteen major tasks were identified in the original Statement of Work as being imperative to the successful completion of this project. These tasks are listed in a table included in Appendix A. Also included is a timeline detailing project progress during Year 03 and plans for a one-year no-cost extension.

Progress in the CQI Function The CQI function involves a research study that is composed of 5 mammography-reading sessions. During each session, a participating radiologist will read a mammographic film and provide an assessment. The radiologist will mark his or her assessments in the CQI software developed specifically for this project and will receive feedback from the program. If the radiologist identifies a malignancy, s/he must indicate on the digitized image on the computer screen where s/he believes the malignancy is located. The first session is considered the "baseline" score for the physician, and the fourth session is considered the follow-up score. Sessions two and three are teaching sessions designed to improve the radiologist's accuracy in reading mammograms. The fifth and final session varies from the first four in that the radiologist will only be allowed to read the digitized image on the computer as opposed to having films available. The purpose of this session is to assess the feasibility of disseminating the CQI over the Internet. Participating radiologists will receive two Continuing Medical Education (CME) credits per session for a total of 10 credits.

During Year 03, project investigators and staff finalized the CQI software. Additionally, all 180 films necessary for the 5 sessions were identified, copied, digitized, and added to the program. The pretest session completed at the end of Year 02 was followed by a pilot session at the beginning of Year 03. Results from the pretest and pilot were used to finalize the film selection and text of each mammographic study for the five sessions. To date, sessions #1 through #3 have been completed. Results of these three sessions are included in Table 1.

Table 1. Results of Sessions 1 – 3 of CQI study.

	Session 1 test session 45 films	Session 2 educational session 30 films	Sessoion 3 educational session 30 films
Average Sensitivity	67%	79%	82%
Average Specificity	82%	74%	77%

Originally, the study protocol called for all five sessions to be completed in Year 03. However, scheduling the radiologists for the sessions took much longer than anticipated. In the original plan, we assumed that the project staff would visit a facility once or twice for each session to administer the program to all participating radiologists working at that facility. Unfortunately, we've discovered that many radiology groups rarely schedule their physicians to be present at the same facility on the same day. Therefore, several trips have been required to complete each session with the radiologists at many facilities. Because these facilities are scattered over all of Washington State and are several hundred miles apart, the scheduling and completion of each session has taken longer than anticipated.

Additionally, the original protocol called for sessions 4 and 5 to be combined and administered during the same visit. Session 4 consists of 45 films and session 5 consists of 30 films for a total of 75 films to be read in the final session. During the earlier sessions, however, we realized that 75 films were too many to ask a radiologist to read in one sitting due to time contstraints and fatigue. Therefore, we have split out sessions 4 and 5. This will add an additional 4 months to the completion of the sessions. We anticipate completing session five during the first half of the no-cost extension.

All five sessions must be complete in order for us to deliver several items promised in our original statement of work. These include evaluating the impact of the CQI on accuracy of interpretation, evaluating inter- and intra- observer variability in film reading, and evaluating the adequacy of using digitized films as a teaching tool.

Progress in the Surveillance and Audit functions The completion of the surveillance and audit functions involve recruiting mammography facilities to participate in the MTR. We consider this a particularly important part of the project as the feedback received from the MTR provides participating facilities and radiologists with accurate assessments of their current practice. Because of the importance of these two functions, the project has focused on continuing to try to recruit those facilities who have yet to sign on to the MTR.

During the course of the last year, most of the 8 MQIP facilities who agreed to participate in the MTR during Year 01 of our project provided us with downloads of their data. This data was then reviewed for errors and cleaned. It was then added to the MTR and linked to cancer data provided by the Washington State Tumor Registry. Table 2 summarizes the data obtained from these facilities. Reports were generated from this link that provided the facilities with summary statistics. Radiologist – specific reports were generated to fulfill the audit function requirements, as well. Examples of these reports are included in Appendix B.

Table 2. Summary of MQIP facilities participating in the Surveillance and Audit functions.

Facility	# mammograms	Years of mammograms
1	51,526	94-98
2	3,890	97-99
3	87,209	91-98
4	9,063	96-99
5	5,775	95-98
6	36,318	95-98
7	615	97-99
8	Data not yet received	Data not yet received

The remaining 19 facilities who are affiliated with radiologists participating in the CQI but who are not members of the MTR were contacted again during Year 03. Many of these facilities have indicated interest in joining at this time. Linking these facilities to the MTR, cleaning their data, and providing reports back to the facilities as well as to the radiologists (the audit function), can take several months, depending on the facility's data system, data quality and their availability to MTR staff. We hope to get at least two additional facilities on board the MTR and receiving reports that satisfy both the surveillance and audit functions of the project within the no-cost extension period. This work will help us to accomplish two other items listed in our original statement of work: linking mammography data to a tumor registry via the MTR, and provide feedback reports to participants.

Progress in the Certification Function Two technologists training conferences were completed during Year 02. These conferences involved specialized sessions developed by professional speakers from the Elizabeth Wende Breast Clinic in Rochester, New York. Local speakers also provided lectures to supplement the conferences.

At the completion of each conference, participating technologists completed a questionnaire asking about the quality of the conference and the information they received while attending it. During Year 03, a follow-up survey was conducted to assess the usefulness of the information received in the conferences six months after they were completed. A final report describing the conferences and their assessment is included as Appendix C. Overall, the conferences provoked a high participation rate, positive reviews, and provided information that was then used in practice after the conference.

The certification function also involves training Washington State Registrars about the importance of accurate TNM staging and rapid reporting of breast cancer cases to the registry. This training benefits the project by improving the quality and timing of data added to the Mammography Tumor Registry (MTR). The MTR is being used by the MQIP to provide surveillance and audit reports for participating facilities.

CTRs began recording TNM staging recently when the American College of Surgeons began to require it for certification for cancer programs in hospitals. TNM is considered to be the most common method of staging and is the standard for the MTR. This staging method is derived from a fairly complex and involved algorithm. Improving the quality of the TNM variable recorded by registrars will reduce the work of MTR staff as well as improve the overall quality of the data.

Quick reporting to the cancer registry is an issue for breast cancer cases, in particular. Because treatment data may come from many sources outside the hospital and may require intensive follow-up by the CTR, it can be difficult to obtain all information in a reasonable amount of time. Reducing the length of time from cancer treatment to entry into the registry ultimately improves the quality of reports generated by the MTR.

To provide training on these two issues, three CTR conferences were conducted in coordination with the Washington State Cancer Registry (WSCR). The conferences were held during Years 01 and 02 of the project. To evaluate the impact of these training sessions, breast cancer cases identified between 1/97 and 12/98 in the WSCR will be analyzed. In particular, we will examine the TNM stage assigned by the registrar and compare it to a gold standard TNM stage established by the MTR. Additionally, we will compare differences in time from hospital admission to registry entry both before and after the completion of the conferences. Results of this evaluation will be provided to the WSCR for feedback to the registrars.

The data necessary to conduct this analysis was to be received from the WSCR in the latter half of Year 03. However, staff turnover in key positions as the WSCR have caused the data to be delayed in delivery. We anticipate receiving the data within the next few months. Evaluation of the registrars training will then be conducted during the no cost extention.

KEY RESEARCH ACCOMPLISHMENTS IN YEAR 03

CQI function

- Conducted pilot
- Finalized the software and session composition
- Conducted sessions 1-3, began session 4

Surveillance and Audit functions

- Linked data from 8 mammography facilities to cancer data
- Generated facility and radiologist specific audit reports
- Conducted recruitment campaign with remaining 19 facilities

Certification function

- Conducted survey to evaluate the success of technologist training
- Generated final report on technologist training
- Prepared for evaluation of registrar training

REPORTABLE OUTCOMES

Manuscripts associated with this project:

- Pepe MS, Urban N, Rutter C, Longton G. Design of a Study to Improve Accuracy in Reading Mammograms. J Clin Epi 1997;50 (12):1327-38 (Included as Appendix D)

CONCLUSIONS

The past year has been very productive for this project. Considerable progress was made in all functions of the MQIP. As part of the CQI function, most of the sessions have been completed and the protocol has been modified to allow for a more reasonable administration of the final two sessions. Recruitment for the surveillance and audit functions has progressed with some facilities showing interesting in joining the MTR at this time. Additionally, facility and radiologist reports have been provided to those facilities who were recruited to the MTR earlier in the project. Evaluation of the registrars training was completed during the last year as part of the certification function. Plans for evaluating the registrars training have been put in place and will be implemented as soon as the data become available.

Because the project is still in the data collection phase, we are unable to write a final report. As is demonstrated in the timeline included in Appendix A, the majority of evaluation activities will be conducted and reported during the no-cost extension.

References

[1] Rosenblatt RA. The Potential of the Academic Medical Center to Shap Policy-oriented Rural Health Research, Academic Medicine 1991;66(11):662-67.

[2] Rosenblatt RA, Lishner DM. Surplus or Shortage? Unraveling the Physician Supply Conundrum. The Western Journal of Medicine, 1991;154(1):43-50.

[3] Brauer GW. Telehealth: the delayed revolution in health care. Medical Progress through Technology 1992; 18:151-63.

[4] Keeler EB, Buinstein LV, Kahn KL, et al. Hospital Characterstics and Quality of Care. JAMA 1992;268:1709-1714.

APPENDIX A

Summary of Major Tasks
Associated with Project and
Detailed Timeline of Project Years 02 and 03

Major tasks listed in original Statement of Work

Function associated with task	Major Task	Progress
All	1. Recruit and enroll radiologists and mammography facilities to MQIP	Radiologist recruitment completed Year 02, facility recruitment ongoing
CQI	2. Obtain CME credit for CQI	Complete, Year 01
CQI	3. Obtain 180 mammograms for 5 sessions of CQI	Complete, Year 03.
CQI	4. Develop software for CQI	Complete, Year 01. Debugging occurred during pretest and pilot in Year 02.
CQI	5. Pilot CQI	Conducted Pretest and Pilot, Year 02
CQI	6. Implement CQI	Initiated Year 03, will complete during extension
Surveillance	7. Develop materials to allow facilities without computerized systems to participate in MTR	Complete, Year 01
Certification	8. Obtain certification for training technologists	Complete, Year 01
Certification	9. Conduct training workshops for technologists	Two training workshops conducted during Year 02.
All	10. Apply for certification of MQIP by Washington State	A federal Certificate of Confidentiality was obtained, Year 02.
All	11. Implement MQIP	Initiated Year 01, will be complete during extension
Surveillance/ Audit	12. Link mammography data to tumor registry via MTR	Initiated Year 02. Ongoing through extension.
Audit	13. Provide feedback reports to participants	Initiated Year 03. Ongoing through extension.
CQI	14. Evaluate impact of CQI on accuracy of interpretation in communities	After implementation of CQI. Will be done during extension.
CQI	15. Evaluate inter-/intra- observer variability	After implementation of CQI. Will be done during extension
CQI	16. Evaluate adequacy of digitized films	After implementation of session 05. Will be done during extension.
Certification	17. Evaluate impact of training CTR's on % of cancer cases entered in tumor registry and quality of data	Data obtained from State Registry last half Year 03. Analysis will be done during extension.
Certification	18. Evaluate usefulness of training program for technologists	Complete, Year 03

Year 03 and extension
timeline for MQIP

Project Year	01	2	03												extension											
Task Month			Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep
CQI																											
Obtain CME accreditation																											
Recruit radiologists																											
Collect films for sessions 1 & 2																											
Develop and fine tune software																											
Pretest CQI																											
Pilot CQI																											
Collect films for sessions 3 & 4																											
Session #1																											
Session #2																											
Session #3																											
Session #4																											
Session #5																											
Evaluate impact of CQI on accuracy																											
Evaluate inter/intra observer variability																											
Evaluate adequacy of digitized films																											
Surveillance and Audit																											
Recruit facilities																											
Develop data entry software for paper facs.																											
Obtain data																											
Link data to tumor registry data																											
Provide reports																											
Certification																											
Obtain certification for tech. conferences																											
First technologist conference																											
Second technologist conference																											
Evaluate usefullness of tech. conferences																											
First registrar's training																											
Second registrar's training																											
Third registrar's training																											
Evaluate impact of training CTR in tumor data																											

APPENDIX B

Manuscript:
Pepe MS, Urban N, Rutter C, Longton G. Design of a Study to Improve Accuracy in Reading Mammograms. J Clin Epi 1997;50 (12):1327-38

Mammography Outcomes Report

Facility: 000
Report Date: 6/30/95

Current Period: 1/1/93 to 12/31/93
Cumulative Period: 1/1/92 to 12/31/93

I. Summary of Mammograms by Exam Type and Assessment

A. Screening (asymptomatic)

ACR		Current		Cumulative	
Code	Description	Exams	Pct of Total	Exams	Pct of Total
0	Incomplete	xx	xx.x	xxx	xx.x
1	Normal	xx	x.x	xxx	xx.x
2	Normal with benign finding	xx	xx.x	xxx	xx.x
3	Probably benign; short term f/u	xx	xx.x	xxx	xx.x
4	Suspicious for malignancy	xx	xx.x	xxx	xx.x
5	Highly suggestive of malignancy	xx	xx.x	xxx	xx.x
	Screening Totals	xxx	xx.x	xxxx	xx.x

B. Diagnostic (symptomatic)

ACR		Current		Cumulative	
Code	Description	Exams	Pct of Total	Exams	Pct of Total
0	Incomplete	xx	xx.x	xxx	xx.x
1	Normal	xx	x.x	xxx	xx.x
2	Normal with benign finding	xx	xx.x	xxx	xx.x
3	Probably benign; short term f/u	xx	xx.x	xxx	xx.x
4	Suspicious for malignancy	xx	xx.x	xxx	xx.x
5	Highly suggestive of malignancy	xx	xx.x	xxx	xx.x
	Diagnostic Totals	xxx	xx.x	xxxx	xx.x
	Overall Totals	xxx		xxxx	

II. Summary of Cancers Associated with Mammograms by Exam Type and Assessment [1]

A. Screening (asymptomatic)

ACR		Current		Cumulative	
Code	Description	Exams	Cancers	Exams	Cancers
0	Incomplete	xxx	xxx	xxxx	xxx
1	Normal	xxx	xxx	xxxx	xxx
2	Normal with benign finding	xxx	xxx	xxxx	xxx
3	Probably benign; short term f/u	xxx	xxx	xxxx	xxx
4	Suspicious for malignancy	xxx	xxx	xxxx	xxx
5	Highly suggestive of malignancy	xxx	xxx	xxxx	xxx
	Screening Totals	xxxx	xxx	xxxx	xxx

B. Diagnostic (symptomatic)

ACR		Current		Cumulative	
Code	Description	Exams	Cancers	Exams	Cancers
0	Incomplete	xxx	xxx	xxxx	xxx
1	Normal	xxx	xxx	xxxx	xxx
2	Normal with benign finding	xxx	xxx	xxxx	xxx
3	Probably benign; short term f/u	xxx	xxx	xxxx	xxx
4	Suspicious for malignancy	xxx	xxx	xxxx	xxx
5	Highly suggestive of malignancy	xxx	xxx	xxxx	xxx
	Diagnostic Totals	xxx	xxx	xxxx	xxx
	Overall Totals	xxx		xxxx	

III. Analysis of Audit Data for Screening Mammograms[2]

	Current	Cumulative
True-Positives (TP)[3]	xx	xxx
False-Negatives (FN)[4]	xx	xxx
True-Negatives (TN)[5]	xx	xxx
False-Positives (FP)[6]	xx	xxx
Positive Predictive Value TP/(TP+FP)	x.xxx	x.xxx
Negative Predictive Value TN/(TN+FN)	x.xxx	x.xxx
Sensitivity TP/(TP+FN)	x.xxx	x.xxx
Specificity TN/(FP+TN)	x.xxx	x.xxx

Mammography Outcomes Report

Notes

1. Breast cancers reported in the Cancer Surveillance System are associated with a mammogram if the cancer is diagnosed within 1 year after the mammographic examination.

2. Mammograms with inconclusive findings (ACR code 0) are excluded from the analysis section.

3. True-Positive (TP): Breast cancer diagnosed within 1 year after a screening mammogram with abnormal findings (ACR codes 4 or 5).

4. False-Negative (FN): Breast cancer diagnosed within 1 year after a screening mammogram with normal findings (ACR codes 1, 2 or 3).

5. True-Negative (TN): No known breast cancer diagnosis within 1 year after a screening mammogram with normal findings (ACR codes 1, 2 or 3).

6. False-Positive (FP): No known breast cancer diagnosis within 1 year after a screening mammogram with abnormal findings (ACR codes 4 or 5).

Facility ID: 999
Physician ID: 99

Detailed Radiologist Feedback Report
(Radiologist Specific)

Report Date: 03/18/99
Date Range: 11/01/91-03/01/96

Patients With Breast Cancers Occurring After a Screening or Diagnostic Mammogram by ACR Assessment

Total mammograms read in period: 915

ACR Assessment: 0 - Needs additional evaluation *Mammograms with this assessment: 21*

Months to Diag[1]	TNM Stage	Exam Date	Diag Date[2]	Laterality[3]	Site[4]	Tumor Size (cm)	DCIS	Vital Status[5]	Rpt Inst[6]	S/D[7]	Patient ID	Patient Name
<1	0	02/14/92	02/92	2	745	0.1	Y	5	CSS	S	000298	DOE, JANE

ACR Assessment: 1 - Negative *Mammograms with this assessment: 755*

Months to Diag[1]	TNM Stage	Exam Date	Diag Date[2]	Laterality[3]	Site[4]	Tumor Size (cm)	DCIS	Vital Status[5]	Rpt Inst[6]	S/D[7]	Patient ID	Patient Name
6	I	12/04/91	05/92	2	744	2.0	N	1	CSS	S	000084	LANE, WILMA
9	IIIA	03/18/92	12/92	1	748	4.0	N	1	CSS	S	000409	SMITH, JULIE
13	0	01/24/95	02/96	1	749	0.2	Y	9	CSS	S	000164	WILLIAMS, MARIE
14	IIA	02/06/92	04/93	1	744	1.8	N	3	CSS	S	000270	WHITE, MARIANN
25	0	01/14/94	02/96	1	749	0.2	Y	9	CSS	S	000164	WILLIAMS, MARIE
36	0	04/30/92	05/95	1	741	0.0	N	3	CSS	S	000575	BROWN, MICHELLE
37	0	01/08/93	02/96	1	749	0.2	Y	9	CSS	S	000164	WILLIAMS, MARIE
38	I	06/10/92	08/95	2	744	1.6	N	3	CSS	S	000058	JOHNSON, EILEEN
41	0	05/09/94	10/97	2	748	0.8	N	2	CSS	S	000262	JOHNSON, BETTY
45	I	04/22/92	02/96	2	744	1.2	N	1	CSS	S	000550	GREEN, JENNIFER
49	0	01/14/92	02/96	1	749	0.2	Y	9	CSS	S	000164	WILLIAMS, MARIE
49	IIB	01/23/92	03/96	2	n/a	5.0	N	1	WSCR	S	000216	JONES, EVA
52	I	02/11/92	06/96	1	744	1.2	N	1	CSS	S	000284	RIVERS, DIANE
64	I	04/01/92	08/97	1	748	0.6	N	1	CSS	S	000462	LAKE, LISA
64	IIIA	04/16/92	09/97	1	748	2.1	N	1	CSS	S	000526	BAKER, CATHY
68	0	02/04/92	10/97	2	748	0.8	N	2	CSS	S	000262	JOHNSON, BETTY

ACR Assessment: 2 - Benign finding - negative *Mammograms with this assessment: 76*

Months to Diag[1]	TNM Stage	Exam Date	Diag Date[2]	Laterality[3]	Site[4]	Tumor Size (cm)	DCIS	Vital Status[5]	Rpt Inst[6]	S/D[7]	Patient ID	Patient Name
28	0	06/13/95	10/97	2	748	0.8	N	2	CSS	S	000262	JOHNSON, BETTY
30	I	07/30/93	02/96	2	744	1.2	N	1	CSS	S	000550	GREEN, JENNIFER

acility ID: 999
hysician ID: 99

Detailed Radiologist Feedback Report (Radiologist Specific)

Report Date: 03/18/99
Date Range: 11/01/91-03/01/96

Patients With Breast Cancers Occurring After a Screening or Diagnostic Mammogram by ACR Assessment

'otal mammograms read in period: 915

CR Assessment: 4 - Suspicious abnormality - biopsy should be *Mammograms with this assessment: 20*

Ionths o Diag[1]	TNM Stage	Exam Date	Diag Date[2]	Laterality[3]	Site[4]	Tumor Size (cm)	DCIS	Vital Status[5]	Rpt Inst[6]	S/D[7]	Patient ID	Patient Name
<1	IIA	03/11/92	03/92	2	743	2.5	N	8	CSS	S	000387	HENDERSON, GLORIA
<1	IIIB	03/09/92	03/92	1	744	5.0	N	5	CSS	S	000379	MCWRIGHT, BETH
<1	IIA	03/30/93	04/93	1	744	1.8	N	3	CSS	S	000270	WHITE, MARIANN

CR Assessment: 5 - Highly suggestive of malignancy *Mammograms with this assessment: 10*

Ionths o Diag[1]	TNM Stage	Exam Date	Diag Date[2]	Laterality[3]	Site[4]	Tumor Size (cm)	DCIS	Vital Status[5]	Rpt Inst[6]	S/D[7]	Patient ID	Patient Name
<1	I	11/08/91	11/91	1	744	1.8	N	1	CSS	S	000038	SMITH, DELORES
<1	I	02/07/96	02/96	2	744	1.2	N	1	CSS	S	000550	GREEN, JENNIFER
1	I	03/23/92	04/92	2	749	1.0	N	3	CSS	S	000420	DERMIT, PAULA
1	I	11/21/91	12/91	1	749	1.2	N	1	CSS	S	000059	O'RILEY, KATHLEEN
1	0	01/17/96	02/96	1	749	0.2	Y	9	CSS	S	000164	WILLIAMS, MARIE
2	I	11/13/91	12/91	2	744	1.1	N	1	CSS	S	000047	KAY, LORRAINE
3	I	11/16/95	02/96	2	744	1.2	N	1	CSS	S	000550	GREEN, JENNIFER
66	IIA	01/17/92	08/97	2	744	3.0	N	1	CSS	S	000188	JACKSON, BARBARA

acility ID: 999
hysician ID: 99

Detailed Radiologist Feedback Report (Radiologist Specific)

Report Date: 03/18/99
Date Range: 11/01/91-03/01/96

Patients With Breast Cancers Occurring After a Screening or Diagnostic Mammogram by ACR Assessment

otal mammograms read in period: 915

NOTES

[1] Estimated number of months from mammogram date to cancer diagnosis. The diagnosis date is estimated by assigning the last day of the month to the diagnosis month reported by CSS.

[2] The CSS registry reports only the month and year of diagnosis.

[3] Laterality: 1 = Right origin of primary
2 = Left origin of primary
3 = Only one side involved, right or left origin unspecified
4 = Bilateral involvement, lateral origin unknown
9 = No information concerning laterality available

[4] Site: 740 = Nipple, Areola
741 = Central portion of breast
742 = Upper-inner quadrant
743 = Lower-inner quadrant
744 = Upper-outer quadrant
745 = Lower-outer quadrant
746 = Axillary tail of breast
748 = Tumor overlaps two or more site boundaries
749 = Breast, not otherwise specified; Breast, mammary gland

[5] Vital Status: 1 = Alive, no clinical evidence or complete remission of cancer
2 = Alive, with any cancer
3 = Alive, cancer status unknown
4 = Deceased, no evidence or complete remission of cancer
5 = Deceased, this cancer present at death
6 = Deceased, no evidence or complete remission of cancer, but another cancer present at death
7 = Deceased, cancer present but not established whether this cancer caused the death
8 = Deceased, unknown whether cancer present at death
9 = Deceased, no follow-up per request of physician and/or hospital cancer registrar

[6] CSS -- Cancer Surveillance System contributing to SEER
WSCR -- Washington State Cancer Registry

[7] S/D: Type of exam taken from facility coding.

'acility ID: 999
'hysician ID: 99

Detailed Radiologist Feedback Report (Radiologist Specific)

Report Date: 03/18/99
Date Range: 11/01/91-03/01/96

Alphabetical Listing of Patient Mammograms Linked to Cancer Diagnoses

'atient D	Patient Name	Exam Date(s)	Dx Date(s)
00526	BAKER, CATHY	04/16/92	09/97
00575	BROWN, MICHELLE	04/30/92	05/95
00420	DERMIT, PAULA	03/23/92	04/92
00298	DOE, JANE	02/14/92	02/92
00550	GREEN, JENNIFER	04/22/92	02/96
		07/30/93	02/96
		11/16/95	02/96
		02/07/96	02/96
00387	HENDERSON, GLORIA	03/11/92	03/92
00188	JACKSON, BARBARA	01/17/92	08/97
00262	JOHNSON, BETTY	02/04/92	10/97
		05/09/94	10/97
		06/13/95	10/97
00058	JOHNSON, EILEEN	06/10/92	08/95
00216	JONES, EVA	01/23/92	03/96
00047	KAY, LORRAINE	11/13/91	12/91
00462	LAKE, LISA	04/01/92	08/97
00084	LANE, WILMA	12/04/91	05/92
00379	MCWRIGHT, BETH	03/09/92	03/92
00059	O'RILEY, KATHLEEN	11/21/91	12/91
00284	RIVERS, DIANE	02/11/92	06/96
00038	SMITH, DELORES	11/08/91	11/91
00409	SMITH, JULIE	03/18/92	12/92
00270	WHITE, MARIANN	02/06/92	04/93
		03/30/93	04/93
00164	WILLIAMS, MARIE	01/14/92	02/96
		01/08/93	02/96
		01/14/94	02/96
		01/24/95	02/96
		01/17/96	02/96

Facility ID: 999
Physician ID: 99

Detailed Radiologist Feedback Report
(Radiologist Specific)

Report Date: 03/18/99
Date Range: 11/01/91-03/01/96

Patients with Negative Mammograms Linked to Cancer Diagnoses Within 24 Months

(includes patients who may have had a positive mammogram following one or more negative mammograms)

Patient ID	Patient Name
000084	LANE, WILMA
000409	SMITH, JULIE
000270	WHITE, MARIANN
000164	WILLIAMS, MARIE

APPENDIX C
Washington Mammography Tumor Registry
Data Flow Overview

The Mammography Quality Improvement Project

Final Report on Training Conferences for Mammography Technologists

Introduction

The Mammography Quality Improvement Project and the Fred Hutchinson Cancer Research Center sponsored two continuing education conferences for mammography technologists in 1998. Each conference was approved by the American Society of Radiologic Technologists (ASRT) for eight Category A credits. There was no fee for conference attendance. Invitations to the conferences were extended to technologists at mammography facilities recruited to or participating in the Mammography Quality Improvement Project.

The first conference, "Picture Perfect", was held on April 25, 1998 at the Hutchinson Center in Seattle. Fifty-three technologists representing 17 facilities attended. The second conference, "Picture Perfect II" was held on October 10, 1998 at Sacred Heart Hospital in Spokane. Fifty-seven technologists representing 14 facilities attended the Spokane conference.

Conference faculty included expert technologists from the Elizabeth Wende Breast Clinic in Rochester, New York, and local radiologists. The overall focus of the conferences was to promote understanding of the breast – its mammographic and physical presentations and patterns of pathology – in order to discern the need and function of supplementary views. Objectives for the conferences were to:

- Develop an appreciation of the anatomical makeup of the breast as it relates to producing the mammogram.
- Review mammography orientation to the breast in terms of tissue summation.
- Review cases with diverse pathology and normal breast patterns.
- Learn to determine when CC and MLO alone are not adequate views for a case.
- Practice positioning.
- Practice problem-solving the necessity for the extra view in some cases.

Focus Group

Two focus groups with mammography technologists were held prior to the conferences to obtain recommendations about how to best design continuing education training for technologists. Specifically our interest was in the educational components that would be most useful to technologists and most likely to improve their clinical experiences. Focus groups were conducted in western and eastern Washington in order to solicit feedback from technologists working in both parts of the state.

There were several common themes from these focus groups. Technologists expressed an interest in learning more about breast pathology, working with live models in "hands-on" positioning workshops, and interacting with technologists from other facilities in a small group setting. There was also a desire to learn how to deal with stress in the workplace, including how to best assist a stressed patient.

Feedback from the focus groups was incorporated into the design and content of the conferences.

Evaluation

The technologist training conferences were evaluated via three surveys: a) on-site survey, b) 2- week post training survey, and c) 6-month and/or1-year post training survey. Complete results of all surveys were provided to conference faculty.

On-Site Survey

The ASRT Education Department requires that sign-in records and course evaluations be submitted no later than 30 days after a training conference. The ASRT provides an on-site evaluation survey that is to be used for each session. Participants may complete the on-site evaluation form for each session immediately following the session, or they may choose to complete all session evaluations at the end of the conference. On-site evaluations were returned to the field coordinators on the day of the conference. The purpose of this evaluation survey was to measure participants' level of satisfaction with the training and to fulfill the requirements of the ASRT.

Each lecture and workshop was rated in five areas: 1) content conducive to learning, 2) manner and skill of presentation, 3) content relevant to work, 4) content covered published course description, and 5) overall satisfaction. Rating selections were "very satisfied", "satisfied", "undecided", "dissatisfied", and "very dissatisfied". There was a space for written comments. The response rates for both conferences were 88% or higher for all sessions.

High levels of satisfaction with all sessions were reported. Results are shown in Table 1 (Seattle) and Table 2 (Spokane). Many participants wrote comments on their evaluations specific to each session. Examples of written comments included: “Great slides, great speaker”, “The room was too dark making it difficult to take notes”, and “Having a live model was good and helped with visualization”.

2-Week Post Training Survey

During the planning stages of the first conference, a consulting technologist suggested that on-site feedback is often rushed, not allowing participants time for thoughtful comments. Therefore, a second evaluation survey was mailed to participants two weeks after the conference. The purpose of the 2-week post training survey was to re-assess participants’ level of satisfaction with the training. In addition to soliciting further comments and criticisms, a second goal was to generate ideas for improving the training format for future conferences.

Overall level of satisfaction was assessed for each session. Response categories were “very satisfied”, “satisfied”, “dissatisfied”, and “very dissatisfied”. To generate ideas for future conferences, qualitative questions regarding interesting and useful topics, potential presenters, training locations, and sponsors were included.

The response rates for this survey were somewhat lower, at 47% for the Seattle training and 56% for the Spokane training. Again, satisfaction levels were high. Results are presented in Table 3 (Seattle) and Table 4 (Spokane).

Many participants provided written responses to the qualitative questions. Examples of written comments included:

- *“I liked the conference because you kept us moving around and the ‘hands-on’ helped me learn.”*
- *“The different stations and variety of lectures kept the day very interesting.”*
- *“This type of conference has a more personal touch to it. It reaches us ‘in the trenches’ so to speak.”*
- *“Have fewer people in the positioning workshop.”*
- *“Have fewer cases and more time with a radiologist in the viewbox workshop.”*
- *“Make it two days and spend more time on positioning and problem solving.”*

6-Month and/or 1-Year Post Training Survey

Participants who attended one or both of the training conferences were mailed a third evaluation survey in March of 1999. The date of this mailing was approximately six months after the Spokane conference and a year after the Seattle conference. The purpose of this survey was to assess the lasting effect of the training and to determine if the skills and knowledge gained at the training contribute to overall quality improvement at the mammography facility. The survey was mailed to 102 technologists with an overall response rate of 57%.

	# who attended	**# surveys returned**	**RR**
Seattle conference only	45	29	64%
Spokane conference only	49	26	53%
Both conferences	8	3	38%
TOTAL	**102**	**58**	**57%**

This survey asked eight "yes" or "no" questions that assessed whether or not the skills and information presented at the conference were used in the technologists' daily routines. One question asked the technologist to choose the one session she found most useful at the conference(s) she attended. The last question on the survey asked the technologist to circle all the reasons she chose to attend the conference(s). Results from the 6-month/1-year post training survey are presented in Tables 5, 6, and 7.

Table 1

Satisfaction Level by Session: Results of the On-Site Evaluation Survey (Seattle)

Session	*% Very Satisfied*	*% Satisfied*	*% Other**
Anatomy and Pathology Lecture			
Content conducive to learning	82	16	2
Manner and skill of presentation	82	14	4
Content relevant to work	78	20	2
Content covered published course description	82	16	2
Overall satisfaction	82	16	2
Problem Solving and Practical Application Lecture			
Content conducive to learning	76	20	4
Manner and skill of presentation	78	16	6
Content relevant to work	82	14	4
Content covered published course description	78	18	4
Overall satisfaction	77	17	6
Pattern Recognition and Pathological Changes Workshop			
Content conducive to learning	43	43	14
Manner and skill of presentation	36	33	31
Content relevant to work	52	36	12
Content covered published course description	45	35	20
Overall satisfaction	37	41	22
Critical Analysis Lecture			
Content conducive to learning	82	18	-
Manner and skill of presentation	84	16	-
Content relevant to work	84	16	-
Content covered published course description	84	14	-
Overall satisfaction	82	18	-
Positioning Workshop			
Content conducive to learning	88	10	2
Manner and skill of presentation	88	12	-
Content relevant to work	88	10	2
Content covered published course description	84	12	4
Overall satisfaction	86	10	4
Problem Solving Workshop			
Content conducive to learning	53	41	6
Manner and skill of presentation	49	33	18
Content relevant to work	65	31	4
Content covered published course description	55	39	6
Overall satisfaction	55	35	10
Nuclear Medicine Lecture			
Content conducive to learning	40	50	10
Manner and skill of presentation	50	44	6
Content relevant to work	31	52	17
Content covered published course description	40	51	9
Overall satisfaction	38	56	6
Applications to Breast Ultrasound Lecture			
Content conducive to learning	68	32	-
Manner and skill of presentation	70	30	-
Content relevant to work	68	28	4
Content covered published course description	68	32	-
Overall satisfaction	68	32	-

***** *Missing, Undecided, Dissatisfied, or Very Dissatisfied***

Table 2

Satisfaction Level by Session: Results of the On-Site Evaluation Survey (Spokane)

Session	*% Very Satisfied*	*% Satisfied*	*% Other**
Anatomy and Pathology Lecture			
Content conducive to learning	90	10	-
Manner and skill of presentation	89	11	-
Content relevant to work	92	8	-
Content covered published course description	87	13	-
Overall satisfaction	87	13	-
Problem Solving and Practical Application Lecture			
Content conducive to learning	71	29	-
Manner and skill of presentation	63	31	6
Content relevant to work	75	25	-
Content covered published course description	69	29	2
Overall satisfaction	67	33	-
Pattern Recognition and Pathological Changes Workshop			
Content conducive to learning	81	19	-
Manner and skill of presentation	77	23	-
Content relevant to work	83	17	-
Content covered published course description	83	17	-
Overall satisfaction	81	19	-
Critical Analysis Lecture			
Content conducive to learning	88	12	-
Manner and skill of presentation	84	16	-
Content relevant to work	92	8	-
Content covered published course description	86	14	-
Overall satisfaction	88	12	-
Positioning Workshop			
Content conducive to learning	86	14	-
Manner and skill of presentation	86	14	-
Content relevant to work	90	10	-
Content covered published course description	88	12	-
Overall satisfaction	86	14	-
Problem Solving Workshop			
Content conducive to learning	83	17	-
Manner and skill of presentation	77	21	2
Content relevant to work	79	21	-
Content covered published course description	81	19	-
Overall satisfaction	75	23	2
Delayed Diagnosis of Breast Malignancies Lecture			
Content conducive to learning	80	20	-
Manner and skill of presentation	80	20	-
Content relevant to work	82	18	-
Content covered published course description	82	18	-
Overall satisfaction	82	18	-

***** *Missing, Undecided, Dissatisfied, or Very Dissatisfied***

Table 3
Overall Satisfaction Level by Session: Results of 2-week Post Training Survey (Seattle)
N= 25

Session	*% Very Satisfied*	*% Satisfied*	*% Other**
Anatomy and Pathology Lecture	88	12	-
Problem Solving and Practical Application Lecture	20	20	-
Pattern Recognition and Pathological Changes Workshop	52	40	8
Critical Analysis Lecture	86	14	-
Positioning Workshop	76	24	-
Problem Solving Workshop	57	43	-
Nuclear Medicine Lecture	17	78	5
Applications to Breast Ultrasound Lecture	61	39	-

*** Missing, Dissatisfied, or Very Dissatisfied***

Table 4

Overall Satisfaction Level by Session: Results of 2-week Post Training Survey (Spokane)

N= 32

Session	***% Very Satisfied***	***% Satisfied***	***% Other****
Anatomy and Pathology Lecture	78	22	-
Problem Solving and Practical Application Lecture	75	25	-
Pattern Recognition and Pathological Changes Workshop	81	19	-
Critical Analysis Lecture	78	22	-
Positioning Workshop	75	25	-
Problem Solving Workshop	81	19	-
Delayed Diagnosis of Breast Malignancies Lecture	69	31	-

***** Missing, Dissatisfied, or Very Dissatisfied***

Table 5
6-Month/1-Year Post Training Survey Questions and Frequencies
N= 58

Survey Question	***Yes n (%)***	***No n (%)***
Has attending one or both of the "Picture Perfect" conferences aided you in your daily routine at work?	58 (100%)	0
Have you utilized any of the positioning techniques that were introduced to you at the conference(s)?	55 (95%)	3 (5%)
Has your understanding of pathology in the breast improved since the conference(s)?	50 (86%)	8 (14%)
As a result of attending the conference(s), do you automatically take additional views to better visualize pathology?	35 (60%)	23 (40%)
As a result of attending the conference(s), do you feel that you have a better understanding of why a radiologist may ask for particular extra views?	51 (88%)	6 (12%)
In your opinion, have your skills as a technologist improved because of what you learned at the conference(s)?	56 (97%)	2 (3%)
After attending the conference(s), did you share any of the information you learned with your colleagues?	54 (93%)	4 (7%)
After attending the conference(s), did you share any of the information you learned with the radiologist(s) at your facility?	37 (64%)	21 (36%)

Table 6
Ratings of Most Useful Conference Session from the 6-Month/1-Year Post Training Survey

Which session did you find most useful at the conference(s) you attended? (circle one)	***n***
Positioning Workshop	18
Problem Solving and Practical Application (viewbox workshop)	13
Pattern Recognition and Pathological Changes Lecture	9
Anatomy and Pathology Lecture	2
Critical Analysis Lecture	1
Missing or more than one answer given	15

Table 7
Reasons for Conference Attendance from the 6-Month/1-Year Post Training Survey

Why did you choose to attend a "Picture Perfect" conference? ***(circle all that apply)***	***n***
No cost	46
Convenient location	41
Topics covered	35
Reputation of presenters	20
I attend as many continuing education conferences as I can	32
I needed CE's	32
Amenities provided	4

APPENDIX D

Pepe MS, Urban N, Rutter C, Longton G. Design of a Study to Improve Accuracy in Reading Mammograms. J Clin Epi 1997;50(12):1327-38

Design of a Study to Improve Accuracy in Reading Mammograms

Margaret Sullivan Pepe,[1] *Nicole Urban,*[1] *Carolyn Rutter,*[2] *and Gary Longton*[1]
[1]DIVISION OF PUBLIC HEALTH SCIENCES, FRED HUTCHINSON CANCER RESEARCH CENTER, SEATTLE, WASHINGTON 98104,
AND [2]CENTER FOR HEALTH STUDIES, GROUP HEALTH COOPERATIVE, SEATTLE, WASHINGTON 98101

ABSTRACT. This paper is concerned with the design and analysis of mammography reading studies. In particular we consider studies aimed at evaluating interventions to improve the accuracy with which mammograms are read. A simple randomized design is suggested in which a relatively large group of readers read sets of mammograms before and after an intervention phase. We propose solutions to three difficult statistical issues which arise in the context of such studies: (i) the choice of primary outcome measure; (ii) the data analysis technique to be employed; and (iii) the methodology for calculating sample sizes for readers and images to be read.

First, we argue in favor of using sensitivity and specificity as the primary outcome measure rather than receiver operating characteristic (ROC) curves, although the latter are considered state of the art for many types of radiology reading studies. We argue that sensitivity and specificity are more clinically relevant and conceptually more straightforward than ROC curves. Second, we suggest a bivariate approach to data analysis for evaluating intervention effects on sensitivity and specificity. This accommodates the correlations inherent between these measures and allows for estimation of joint effects on them. Finally we propose a method for power calculations which utilizes computer simulation techniques. Simple formulas for sample size calculations are not available in part because variability in accuracy amongst readers and variation in difficultly amongst images introduce complexity into power calculations. The simulation method which we propose accommodates such complexity and is easy to implement.

The methodology was motivated by a study funded by the Department of Defense to evaluate the potential efficacy of an educational intervention. In the context of this study we illustrate the steps involved in power calculations and apply the data analytic techniques to the sort of data expected to result from this study. Though the proposed methods were motivated by this particular study, the statistical considerations are relevant more broadly in mammography and indeed in other types of radiologic imaging studies. Standards for the conduct of radiologic reading studies are not yet well developed, as they are for randomized clinical trials and for case-control studies. We hope that the discussion in this paper will add to the dialogue necessary for development of such standards.

KEY WORDS. ROC curves, sensitivity and specificity, computer simulation, diagnostic tests, screening.

1. INTRODUCTION

Mammography screening for breast cancer has been shown to be associated with decreased breast cancer mortality, at least in women over the age of 50 years [1]. Major efforts are currently underway to improve participation by women in screening programs [2]. Nevertheless, there is concern about the quality of mammography screening and there is general agreement that improvements in quality may lead to improvements in the performance of mammography as a screening modality. Quality might be improved for example by improving the imaging procedures. Alternatively, improvements in the accuracy with which mammographers interpret mammograms may improve the performance of screening mammography. Recent studies [3,4] have shown that there is considerable variability amongst radiologists in their interpretations of screening mammograms. Elmore et al [3] observed that sensitivities ranged from 74% to 96% and that specificities ranged from 35% to 89% among 10 radiologists reading 150 selected mammograms. Beam et al [4] using a much larger sample of 108 radiologists, each reading 79 mammograms, found sensitivities in the range of 47% to 100% and specificities in the range of 35% to 99%. These observations suggest that improvement in interpretation may be possible.

As part of a project called the Mammography Quality Improvement Project (MQIP) funded by the Department of Defense and aimed at improving the quality of mammography screening in rural communities, we are developing an educational program to improve the accuracy with which radiologists interpret mammograms. The educational intervention is comprised of a series of five sessions in which mammographers read films and are provided with immediate feedback on the accuracy of their interpretations. Feedback is provided using a laptop personal computer which is mailed to the radiologist prior to his reading session. The computer program emphasizes the particular features of each mammogram which are relevant to determining the disease status of the woman screened. Eventually it may be possible to disseminate this sort of intervention over computer networks thus making it attractive in terms of easy accessibility and low cost.

In order to evaluate the impact of such an intervention on improvements in diagnostic accuracy it will eventually be necessary to perform a study of radiologists interpretations of screening mammograms in their actual practices. As a preliminary step to such a large-scale study, we will evaluate the intervention effects in a more controlled setting. Specifically, we will have a number of radiologists read a selected set of mammograms before and after the intervention and evaluate changes in accuracy. The mammograms included in this controlled study will be comprised of about 50% from women with disease, a proportion much larger than would be observed in practice but necessarily high in order to estimate sensitivity rates in a small-scale study. Mammograms will be selected to represent a reasonably broad range of interpretive difficulty.

The purpose of this paper is to elucidate some of the key statistical issues in the design of such a controlled reading study. Standards for the design of such studies are not well developed. This contrasts with therapeutic clinical trials and epidemiologic studies where the basic elements of study design are now fairly well standardized [5]. The question we propose to address in this reading study, namely evaluation of an intervention effect in a controlled setting, is a standard sort of question addressed in diagnostic imaging research. Hence the design issues which are dealt with here will have implications for future studies in mammography and in other diagnostic test settings. These same issues also arise in reading studies designed to compare different imaging modalities. The key issues concern the choice of relevant primary outcome measures, appropriate data analysis strategies, and methodology for power calculations which incorporates variability amongst radiologists and amongst images. Broader issues in regards to study designs for evaluating imaging tests have been discussed in a more general sense in the literature [6,7].

In Section 2 we consider two sets of measures which can be used to define accuracy in reading mammograms; firstly, sensitivity and specificity and secondly, ROC curves. We argue in favor of the former, in part, because they are more clinically relevant and most easily understood but also because the latter can provide inappropriate conclusions concerning intervention benefits. In Section 3 we detail the basic elements of the statistical design of our study which could be considered a prototype for evaluating intervention effects in diagnostic radiology. An approach to joint analysis of sensitivity and specificity is outlined in Section 4. In Section 5 we describe methodology for power calculations which are appropriate for the proposed design and analysis. We propose the use of computer simulation methods for calculating power because they allow for complex designs and can easily incorporate variability amongst radiologists and images. Having described the steps involved in calculating power in Section 5, we then apply these procedures to the proposed MQIP study in Section 6, in order to illustrate the methods. Concluding remarks follow in Section 7.

2. MEASURES OF ACCURACY

2.1 Definitions

A radiologist reading a set of mammograms for a woman in our study will classify each breast according to his or her suspicion of its showing malignancy. The ACR lexicon for rating a breast [8] which we will employ, defines a 5-point scale with category 1 indicating "normal, routine follow-up recommended", 2 indicating "benign, routine follow-up", 3 indicating "probably benign, early recall recommended", 4 indicating "suspicious for cancer, consider biopsy" and 5 indicating "highly suspicious for cancer, biopsy recommended". A common definition of a screen positive mammogram is one which receives a rating of 4 or greater. These are mammograms which are sufficiently suspicious for cancer that biopsy is recommended and hence they have an impact on clinical practice. Sometimes a rating of 3 or greater is considered positive. Because of the clinical implications of ratings 4 and 5, we will focus on the positivity criterion of category ≥ 4 here.

Given a definition for screen positivity, since there is a rating for each breast, one can calculate sensitivities and specificities with either 'woman' or 'breast' as the unit of analysis. The latter includes all non-diseased breasts, (including non-diseased breasts from women with cancer), as the denominator for specificity and all diseased breasts as the denominator for sensitivity. Breast level definitions are arguably inappropriate in clinical studies of mammography. It seems more clinically relevant to use woman as the unit of analysis. For this, one could use the maximum of the ratings for the left and right sides as the woman level rating for calculation of sensitivity and specificity. However, occasionally a woman with unilateral disease may not have it detected in the affected side but will have a positive mammogram on the unaffected side. In this case, using the maximum rating will inappropriately inflate the sensitivity. We define sensitivity instead as the proportion of women with disease who have it detected (a rating of ≥ 4) on the affected side. The specificity is the proportion of women without disease who have a maximum rating of less than 4.

ROC analysis is a statistical technique used to describe accuracy of diagnostic tests when the test outcome is either ordinal or continuous as opposed to binary. The rating data generated in radiology reading studies is ordinal and ROC analysis is often considered optimal for the analysis of such studies as is evidenced, for example, in a recent issue of *Academic Radiology* [9]. An ROC curve is constructed by varying the criterion used for defining a positive mammogram from "rating ≥ 2" to "rating ≥ 5", plotting the associated sensitivity and 1-specificity values against each other, and finally fitting a curve to the points so that the curve is anchored at (0,0) and (1,1). Various algorithms exist for fitting a curve, the most notable being the Dorfman-Alf algorithm based on the binormal model [10] and the empirical nonparametric method which simply connects observed ROC points linearly. The area under the ROC curve is usually used to summarize accuracy. Again we suggest that woman rather than breast should be the unit of analysis in defining the ROC curve. That is, in calculating the sensitivity corresponding to the criterion "rating $\geq K$", it should be defined as the proportion of women with cancer who have a rating of $\geq K$ on an affected side.

2.2 ROC Analysis Versus Sensitivity and Specificity

ROC analysis was developed originally for diagnostic tests with results on some arbitrary scale. Its primary advantage is that it allows one to assess the inherent capacity of the test to distinguish between diseased and non-diseased subjects without linking the test to some particular threshold for defining screen positive [11,12]. This seems appropriate in radiology experiments when image ratings are arbitrary numbers with no specific clinical meaning attached to them. In that case, shifts in the distributions of ratings are of no consequence as long as they are equally shifted for diseased and non-diseased subjects. In mammography, however, mammogram ratings have very specific clinical meanings and consequent clinical implications. Uniform shifts in the frequencies with which rating categories are chosen can have major clinical implications.

Moreover, in contrast to the prototype setting for ROC analysis, shifts between certain diagnostic categories are of more importance than others. For example, as noted by Kopans [13], whether an image is rated in category 4 versus category 5 has no clinical impact. Similarly classifications in category 1 versus category 2 are clinically irrelevant. However, shifts between categories 4 or 5 and between 1 or 2 can have a big impact on the ROC analysis. To illustrate this consider the setting shown in Figure 1. The effect of intervention in this setting is to shift classifications of diseased observations from category 4 to category 5 and classification of non-diseased subjects from category 2 to category 1. Though these changes are of no clinical import, the ROC type analysis indicates a benefit for the intervention. Thus an ROC analysis can indicate a benefit of intervention even though a clinically relevant benefit does not exist.

Of even more concern is the fact that a clinically relevant benefit of intervention can occur even when the ROC curves pre- and post-intervention are the same. Consider the ROC curve depicted in Figure 2 for such a situation. The location on the ROC curve of the points associated with the criterion "rating $\geq$ category 4" indicate that sensitivity was significantly increased without decreasing specificity. This clinically relevant improvement in test accuracy does not manifest itself in an improvement in the ROC curves since the pre-and post-intervention curves are the same. (Interestingly, classic binormal ROC curves do not fit the situation depicted in Figure 2 and a binormal ROC analysis in this setting may incorrectly indicate that the ROC curve post-intervention is improved over that pre-intervention).

The fact that ROC analysis can yield inappropriate conclusions regarding the clinically relevant effects of intervention argues against its use for the primary analysis of mammography reading study data. Another valid argument for not using an ROC analysis is that it is complicated and not easily understood by clinicians. Moreover, the so-called 'area under the curve' which summarizes the ROC curve in a single number has an interpretation which is not well known or easily understood. It can be interpreted as the probability that a radiologist will have a greater suspicion of cancer from a mammogram from a woman with disease than from a woman without [14]. This probability, however, seems to be of more theoretical than practical relevance.

We propose using the more clinically meaningful quantities of sensitivity and specificity for the primary data analysis and employing ROC analysis as a secondary descriptive device. Though ROC analysis may be statistically more powerful in some settings, any study should be designed so that it has adequate power to detect changes in the quantities which are of practical relevance. Hence, we suggest that power calculations for a mammography reading study should be based on the ability to detect changes in sensitivity and specificity rather than on the basis of detecting changes in ROC curves.

3. STUDY DESIGN

We now describe the basic elements of the design which we propose for studies evaluating intervention effects on reading accuracy in mammography. In this prototype design, radiologists are randomly assigned to intervention and control groups, with the number in the former being denoted by R_T and the number in the latter denoted by R_C. Two image sets are constructed with M images in each set $S = 1, 2$. In set S, a number M_D^S are from women with disease and this number may differ between the two sets. Each reader reads one set of images before the intervention period and one set after. It is important that the sets before and after intervention be different since readers may remember, to some degree, images that they have previously read. Half of the readers chosen at random in each of the intervention and control groups read set 1 before intervention and set 2 after intervention. The other half read them in the opposite order: set 2 followed by set 1. This cross-over of film sets eliminates the possibility of systematic bias due to film sets. The design is balanced in the sense that set 1 is read equally often before and after the intervention phase in both the intervention and control groups, and similarly for set 2. Readers are told the approximate prevalence of diseased images, i.e., $(M_D^1 + M_D^2)/2M$ and that this varies between the two sets. The rationale for telling the readers the approximate prevalence is that it will become apparent in any case after reading the first set of images and that *apriori* knowledge of it should reduce the potential impact as much as possible on the observed improvement in accuracy. Readers will use the ACR lexicon to classify mammograms and for each reading it will be determined if it is screen positive or negative according to whether the rating is greater than 3 or not.

Images for inclusion in the study need to be selected so that average sensitivity and specificity at the baseline assessment are relatively low. That is, improvements in accuracy should be possible with the sets of images chosen. If, in the absence of intervention all images from women with disease were easily identified as such, the observed sensitivities pre- and post-intervention would be close to 1 and a change in sensitivity would not be identifiable regardless of the actual effect of intervention. Thus at least some of the diseased images should be difficult but not impossible to identify as being from women with disease. Analogous considerations apply to specificity and the choice of non-diseased images included in the study.

4. DATA ANALYSIS

Having described the basic elements of the design and the choice of primary outcomes, we turn now to the strategy for data analysis. There are two components to the analysis. The first concerns a comparison of post- versus pre-intervention reading accuracy among the R_T readers in the intervention group. The second is the comparison of changes from pre- to post-intervention between the intervention and control groups. We first consider the former analysis, in part because it allows us to define notation most easily.

The purpose of this data analysis is to compare the overall sensitivity pre-intervention with that post-intervention and to compare the overall specificity pre-intervention with that post-intervention. If $\hat{S}_{r,\mathrm{pre}}$ and $\hat{S}_{r,\mathrm{post}}$ denote the observed pre- and post-intervention sensitivities for radiologist r, then the observed change in the overall sensitivity $\hat{\Delta}$(sensitivity) is the average change in sensitivities across radiologists in the intervention group:

$$\hat{\Delta}_T(\text{sensitivity}) = \frac{1}{R_T}\sum_{r=1}^{R_T}\left(\hat{S}_{r,\mathrm{post}} - \hat{S}_{r,\mathrm{pre}}\right).$$

Similarly the observed change in the overall specificity in the intervention group is

$$\hat{\Delta}_T(\text{specificity}) = \frac{1}{R_T}\sum_{r=1}^{R_T}\left(\hat{F}_{r,\mathrm{post}} - \hat{F}_{r,\mathrm{pre}}\right)$$

where $\hat{F}_{r,\text{pre}}$ and $\hat{F}_{r,\text{post}}$ denote the observed pre- and post-intervention specificities for radiologist r. Variance estimators for $\hat{\Delta}_T$(sensitivity) and $\hat{\Delta}_T$(specificity) are provided in the appendix. Although $\hat{\Delta}_T$(sensitivity) and $\hat{\Delta}_T$ (specificity) are sample means of changes in sensitivities and specificities, their variances are not given by the usual variance formulae for sample means. Indeed such sample variances would overestimate the variability. Rather the correct variance estimators rely on acknowledging that there are in essence two strata of radiologists in the design, which are defined by the ordering of the two image sets which are rated. The variances of $\hat{\Delta}_T$(sensitivity) and $\hat{\Delta}_T$(specificity) are averages of stratum-specific variances, as shown in the appendix.

Sensitivity and specificity are highly correlated parameters. Radiologists with high sensitivities tend to have low specificities. This will happen for example if they have a low threshold for classifying images as diseased. Similarly, changes in sensitivities and specificities induced by the intervention may be highly correlated. In particular, if the intervention simply changes the implicit threshold a radiologist has for classifying a mammogram as diseased then the sensitivity and specificity will both be changed but in opposite directions. Thus it is important to assess joint effects of intervention on sensitivity and specificity and to account for correlations between them in making inference. This can be accomplished by employing a bivariate analysis approach which is a special case of multivariate analysis, and for which there is a large statistical literature [15]. Using this approach to test the hypotheses that the true average sensitivity and specificity are unchanged by the intervention, $H_0 : \Delta(\text{sensitivity}) = \Delta(\text{specificity}) = 0$, a chi-squared test statistic is calculated. This statistic is a function of the observed average changes, $\hat{\Delta}$(sensitivity) and $\hat{\Delta}$(specificity), their variances and also their correlation. An expression for the chi-squared statistic is provided in the appendix.

In addition to simply testing the hypothesis of no intervention effect, it will be important to provide a confidence region for the intervention effects on sensitivity and specificity based on the observed data. That is, a range of intervention effects, $\{\Delta(\text{sensitivity}), \Delta(\text{specificity})\}$, which are consistent with the observed data. Such a joint 95% confidence region is defined formally as the set of values (x, y) for which the hypothesis H_0 : $\{\Delta(\text{sensitivity}) = x, \Delta(\text{specificity}) = y\}$ is not rejected at the 5% significance level. This region is an elipse, centered at the observed intervention effect ($\hat{\Delta}$(sensitivity), $\hat{\Delta}$(specificity)). We refer the interested reader to the text [15] by Johnson and Wichern (1988, section 5.2) for technical details regarding its calculation. Code for calculating such regions has been written by Murdoch and Chow for the S-PLUS statistical software package and can be obtained from the S-archive on the Statlib computer site (http://lib.stat.cmu.edu). In a similar fashion a joint confidence region for the overall average sensitivity and specificity pre- or post-intervention can be calculated. It is calculated using the observed radiologist specific sensitivities and specificities pre- and post-intervention, and requires only calculation of the means, variances and correlations for these parameters. In order to illustrate these analyses, Figure 3 displays joint confidence regions based on a simulated data set. In our opinion these confidence regions provide a simple summary of the information contained in study data regarding intervention effects on reading accuracy. In the simulated data, the analyses show that sensitivity was increased by the intervention whereas there is no evidence of change in specificity.

So far we have considered the comparison of post- versus pre-intervention reading accuracy within the intervention group. In order to attribute changes in accuracy to the intervention it will be necessary to compare the changes in the intervention group with those in the control group. Without the control group comparison observed changes might be attributed to other factors, such as the increased reading practice or increased awareness of reader fallibility induced by participation in the study. Thus, turning now to the comparison of intervention and control groups, the main hypothesis to be tested is that the changes in sensitivity and specificity in the intervention group are the same as those in the control group. Using a subscript T to denote the intervention group and subscript C to denote the control group, the null hypothesis is $H :_0 \ \Delta_C(\text{sensitivity}) = \Delta_T(\text{sensitivity}), \Delta_C(\text{specificity}) = \Delta_T(\text{specificity})$. A test statistic which has a chi-squared distribution with 2 degrees of freedom is described in the appendix for testing this hypothesis. Joint confidence regions for the differences in changes between the groups, namely $\Delta_T(\text{sensitivity}) - \Delta_C(\text{sensitivity})$ and $\Delta_T(\text{specificity}) - \Delta_C(\text{specificity})$, can be calculated using methods analogous to those described earlier for the pre-versus-post-intervention comparison.

5. METHODOLOGY FOR POWER CALCULATIONS

Power calculations for the reading study are somewhat complicated. They must accommodate the facts that readers vary in their accuracy parameters of sensitivity and specificity, that their sensitivities and specificities are likely negatively correlated, that images vary in difficulty and that a bivariate analysis approach will be employed. These factors together make analytic expressions for sample size intractable. We instead take a computer simulation approach to power calculations. The simulation approach to power calculation is a general and standard method and indeed software has been developed for certain types of applications [16]. The basic idea is to repeatedly simulate data as it is expected or hoped to arise in the course of the study, and determine how often the null hypothesis is rejected. By definition the statistical power of the study is the proportion of simulated studies in which the null hypothesis is rejected. One calculates the power in this fashion using various sample sizes until a sample size is found which provides adequate power. This indirect computer intensive approach to sample size calculation is easily accomplished with modern computers.

5.1 Models for Pre- and Post-intervention Accuracy

In order to simulate study data we need to define precisely the mechanisms giving rise to the data. We therefore need to make assumptions about the reading accuracies before and after intervention. For this purpose we suppose that before intervention a reader correctly assesses a woman with tumor as being diseased with probability $P^D_{r,i}$. The probability $P^D_{r,i}$ depends on the image denoted by i and on the reader, denoted by r. The probabilities $P^D_{r,i}$ will presumably be higher if the tumor is clearly visible in image i than if it is not. The probabilities will also be higher if the radiologist is conservative and is inclined to recommend biopsy for borderline cases. We let S^D be the sensitivity of the average radiologist to the average film from a woman with tumor. The variability amongst films in terms of the difficulty which readers have in assessing them, is captured by specifying a distribution for the sensitivities that the average reader has in assessing the films. Here we assume that the average reader's sensitivity to films varies uniformly in an interval $(S^D - a^D, S^D + a^D)$ across different films. Thus for the average radiologist, easier films are read with sensitivity closer to $S^D + a^D$ and more difficult films are read with sensitivity closer to $S^D - a^D$. In a similar fashion, on the average film from a diseased woman, the sensitivity of different readers is assumed to vary uniformly in an interval $(S^D - b^D, S^D + b^D)$

across radiologists. Thus radiologists with high sensitivity to the average film will have sensitivity closer to $S^D + b^D$. In the appendix we detail a logistic model with random effects (also called a mixed model) for the probabilities $P^D_{r,i}$ which give rise to inter-image and inter-reader variability as postulated here. It is assumed that on the logistic scale there are no interactions between reader and image specific effects on the sensitivity.

Observe that for the purposes of simulating data, by specifying S^D and a^D we can now generate a random image effect by choosing a random number in $(S^D \pm a^D)$ which corresponds to the sensitivity an average radiologist has for detecting it. Similarly, having a specified S^D and b^D we are in a position to generate a random reader effect by choosing a random number in $(S^D - b^D, S^D + b^D)$ which corresponds to his sensitivity to the average film. The logistic model displayed in the appendix then yields the probability $P_{i,r}$ which that reader has of correctly assessing that image as diseased.

Analogous considerations apply to the determination of randomly generated specificities which vary across radiologists and across images from women without disease. Values for parameters $S^{\bar{D}}$, $b^{\bar{D}}$ and $a^{\bar{D}}$ need to be specified in order to define the data generating process. Here, $S^{\bar{D}}$ is the probability that the average radiologist will correctly assess the average non-diseased image as such, radiologists vary uniformly in $(S^{\bar{D}} - b^{\bar{D}}, S^{\bar{D}} + b^{\bar{D}})$ in their specificities to the average non-diseased film, and images from women without disease vary uniformly in $(S^{\bar{D}} - a^{\bar{D}}, S^{\bar{D}} + a^{\bar{D}})$ in the probabilities of the average reader correctly classifying them. The sensitivities and specificities from single radiologists should be correlated. In the appendix we describe how negative correlation between sensitivities and specificities within radiologists can be built into the data simulation mechanism.

In summary, for each study radiologist we simulate his/her sensitivity and specificity to the average diseased and non-diseased films respectively, by randomly sampling correlated numbers from $(S^D - b^D, S^D + b^D)$ and $(S^{\bar{D}} - b^{\bar{D}}, S^{\bar{D}} + b^{\bar{D}})$, respectively. For each study film we determine the sensitivity or specificity that an average radiologist has for it by randomly sampling a number from $S^D - a^D, S^D + a^D)$ or $(S^{\bar{D}} - a^{\bar{D}}, S^{\bar{D}} + a^{\bar{D}})$. Finally, for each combination of film i and radiologist r, we can calculate $P^D_{i,r}$ or $P^{\bar{D}}_{i,r}$, which is the probability that the radiologist will assess that image correctly.

The $P^D_{i,r}$ and $P^{\bar{D}}_{i,r}$ pertain to probabilities before intervention in the treatment and control groups. One also needs to specify treatment effects in order that corresponding probabilities after intervention can be calculated. We postulate that after intervention the quantities S^D and $S^{\bar{D}}$ are changed to new values but that the variations amongst readers and amongst images remain the same. In the appendix we define in a mathematically precise way a logistic model which incorporates such intervention effects.

5.2 Simulated Study Data Generation

Having specified statistical models for pre- and post-intervention rating probabilities which incorporate variation amongst radiologists and amongst images, we now turn to the simulation of study data in accordance with the study design which we proposed in section 3. The first step is to generate images and image sets. This entails generating M diseased images (i.e., M image-specific parameters, one for each image), generating M non-diseased images, and finally from the $2M$ films choosing M at random without replacement to form film set 1. The remaining M films constitute film set 2. The next step is to generate R_T intervention readers and R_C control readers and assign them film sets. That is, for each of $R_T + R_C$ readers we generate pairs of pre- and post-intervention sensitivities and specificities to average diseased and non-diseased films according to the models described in section 5.1. Of the total $R_T + R_C$ readers, R_T are assigned at random to the intervention group and the remaining R_C to the control group. Finally film set orderings are assigned to the readers with half of the intervention readers selected at random being assigned set 1 first and the other half assigned set 2 first. Similarly, $R_C/2$ control readers are assigned set 1 followed by set 2 and the other $R_C/2$ readers are assigned film sets in the opposite order.

The final step in generating data for a simulated study is to actually generate the readings for each reader and image combination. That is, for each reader and for each of the M films in his/her pre-intervention set, a binary random variable is generated which is his/her assessment of whether or not that image shows disease using the probability $P^{D}_{r,i,\text{pre}}$ if the image is diseased and $1 - P^{\bar{D}}_{r,i,\text{pre}}$ if the image is not diseased. Similarly, for each of the M films in his/her post-intervention set a similar binary random variable is generated using $P^{D}_{r,i,\text{post}}$ or $1 - P^{\bar{D}}_{r,i,\text{post}}$ noting that the pre- and post-probabilities differ by different amounts for intervention-versus-control radiologists.

Having generated the simulated study data the test statistics of interest can now be calculated. Data are simulated (first the probabilities, then the ratings) and results calculated under the same assumptions and study design many times, with 1000 or 5000 simulated datasets being typical numbers used for power calculations. The proportion of simulated studies in which the null hypothesis is rejected is the calculated study power for that design and under those assumptions.

6. POWER CALCULATIONS: RESULTS FOR THE MQIP STUDY

To fix ideas, we now illustrate the computer simulation method for power calculations in the MQIP study. This illustration also identifies some sources of data to guide assumptions for power calculations.

We need to choose assumed parameters for the baseline sensitivities and specificities, for the variations amongst radiologists and amongst images and for intervention effects of interest. We assume that the median sensitivity pre-intervention, S^D, in our study will be in the range of .70 to .80. This accords with previous studies which found median sensitivities of .70 and .80 [3,4]. Median pre-intervention specificity will also be assumed to lie in the range of .70 to .80. Beam et al [4] found a median specificity of 0.94 for mammograms from women with normal mammograms and a median specificity of 0.60 for mammograms from women with benign disease. Elmore et al [3] found a median specificity of 0.94. In contrast to these studies we will inform the radiologists of the average prevalence which is higher than that expected in a practical screening setting. Because of this and the fact that the films in our study will be somewhat difficult, we anticipate an initial specificity lower than observed in those studies. The variation amongst radiologists in sensitivities and specificities will be assumed such that $b^D = 0.20$ and $b^{\bar{D}} = 0.20$, which is in agreement with the range of approximately 40% in sensitivities (and specificities) amongst radiologists observed in Beam's study. We could find no data on inter-image variability to suggest appropriate values for a^D and $a^{\bar{D}}$. We assume that they are of the same order of magnitude as the inter-rater variability parameters, $a^D = a^{\bar{D}} = .20$. In regards to intervention effects of interest, we consider that changes of 10 percentage points in either sensitivity or specificity are of interest. However, we calculated power for a variety of intervention effects.

Practical considerations concerning time and cost dictate the range of sample sizes which are feasible and therefore, for which power calculations are performed. We anticipate that no more than approximately 80 radiologists are available for the reading study in the rural communities in which our mammography quality improvement study is being conducted. To maximize power, equal numbers of radiologists are assigned to control and intervention groups. Therefore the number of radiologists per group to be considered for power calculation purposes will be in the range of 20 to 40. Experience suggests that readers can comfortably read no more than 45 films per session. We therefore calculated power for experiments in which the number of films per set, M, was either 30 or 45.

Estimates of power based on computer simulations are shown in Table 1. Though results are shown only for intervention effects on sensitivity with no effect on specificity, because of the symmetry inherent in the design, the same power calculations hold for a 10% change in specificity with no change in the sensitivity. Observe that the power is far larger for the within intervention group assessment of change than for the between group comparison of change. This is to be expected since the variability involved in comparing two random changes is greater than the variability involved in comparing a single change with the null hypothesis of no change. We also observe from Table 1 that the power is less when the baseline sensitivity is 0.70 than when it is 0.80. This is due to the relatively larger binomial variance for the lower baseline rate. In order to be conservative we focus on this lower rate. Interestingly the baseline specificity had little impact on the power to detect an intervention effect on the sensitivity.

The target power for our study design is 90% which allows a 10% chance of an inconclusive result when the intervention increases sensitivity from .70 to .80. For the within intervention group comparison this cannot be achieved with 20 readers but can be achieved with 30 readers if 45 images are included in each image set. The between group comparison, however, has a power of only 66% in this case. Even with use of our maximum resources, i.e., 40 readers per group and 45 images per reading set, the power is only 80%. This allows for a 20% chance of an inconclusive result even when there is a clinically important intervention effect on diagnostic accuracy.

For the MQIP study we chose to focus the study on the within group comparison. The power calculations were an important contribution to this decision but other considerations also played a role. Radiologists would have little motivation to participate in the control arm whereas they would receive CME credit for participation in the intervention arm. The possibility that those in the control arm would learn from the baseline assessment was also a concern and thus we were concerned that it might not even be feasible to construct a true control group. Finally, it was felt that if we found a definite positive change in the intervention group, then this would provide sufficient motivation to proceed with more comprehensive controlled studies in the future. Thus we chose to study only the intervention effects in the intervention group and to use sample sizes of 30 radiologists each reading sets of mammograms from 45 women before and after intervention.

The simulation program allowed us the flexibility to explore the performance of this study design in a variety of settings other than that assumed for the primary sample size calculation. First we calculated the probability of rejecting the null hypothesis for settings where there was no intervention effect. Recall that inference for the test statistic is based on a chi-squared statistic and is theoretically valid with large samples. However, this study entails relatively small samples. We used the simulations to check the adequacy of the large sample theory in our study. To do this we generated data under the null hypothesis. The rejection probability was approximately .06 in the settings we studied, indicating that the true significance level of the test is slightly higher than the target of .05 but adequate for our purposes.

We next explored the power of this study design and sample sizes to detect an array of intervention effects. Results are shown in Table 2. Although the study has adequate power to detect a change in sensitivity (or specificity) of 0.10 even when the pre-intervention sensitivity is as low as 0.60, it has little chance of detecting a smaller change of 0.05. On the other hand, if small changes of the order of .05 occur in both the average sensitivity and in the average specificity there is a good chance that the simultaneous effects will be detected.

7. DISCUSSION

Diagnostic imaging technology is already a basic component of medical care and continues to develop at a rapid pace. It is clearly important to assess the accuracy with which readers can diagnose disease using such technologies, to evaluate the effects of training strategies and to compare methods. Implications for public health can be enormous. Unfortunately, statistical methodology for evaluating and comparing imaging methods has not received much attention by biostatisticians and epidemiologists involved in public health research. Rather the literature is concentrated in radiology research journals, has generally focused on small scale studies involving only a few readers and has ignored clinical implications associated with different diagnostic categories. We believe that it is time to bring the discussion about study design and analysis for evaluating imaging technology to the broader community of epidemiologists and statisticians involved in public health. This is particularly important as interest increases in the accuracies and costs of these imaging methods. By presenting our thoughts on the design and analysis of a study to evaluate an educational intervention on the interpretation of mammograms, we hope to stimulate such discussion.

The choice of primary outcome measure is the most basic element of any study design. We chose to consider the sensitivity and specificity as the basis for evaluating intervention effects. This conflicts with initial statistical reviewers of our study design who were of the opinion that ROC analysis was the only appropriate and indeed the state-of-the-art basis for evaluating an intervention effect. We now argue that in mammography where specific clinical actions are associated with diagnostic rating categories, sensitivity and specificity provide a more clinically relevant and conceptually straightforward basis for comparison than does ROC analysis. Moreover this approach allows us to evaluate effects on false positive as well as true positive rates. In contrast ROC analysis does not quantify the false positive rates directly but in a sense only uses it to standardize the true positive rate. We do not dismiss ROC analysis entirely but rather we regard the analysis of the specific rating categories of secondary importance and focus the design on sensitivity and specificity. Thus the MQIP study was designed to ensure adequate power to detect changes in the most clinically relevant quantities.

We also needed to decide upon the analysis techniques for making statistical inference about sensitivity and specificity. We propose to simultaneously estimate sensitivity and specificity using multivariate methods. Sensitivity and specificity as we have defined them are average sensitivities and average specificities of radiologists in our study. They can also be interpreted as marginal or population average quantities, in the sense of being the probability that a diseased (or non-diseased) image will be correctly interpreted as such in the study. The distinction between the population average and average radiologist-specific interpretations has to do with whether one considers the accuracy parameters to be based on data pooled across radiologists (population average) or to be based on calculation of the accuracy parameter for each radiologist and then averaging the results. In our study these quantities coincide because all radiologists expect to read the same numbers of films. In studies where this is not the case, the distinction should be considered and a decision should be made regarding which of the two entities is most relevant.

The approach we propose for statistical inference is relatively straightforward, being based on methods for inference about sample means. Confidence intervals are based on the variance-covariance matrix of the estimated (sensitivity, specificity) parameters or their changes amongst radiologists. Possible non-normality of the average estimates may be an issue in our study, though for the settings considered in the power calculation this did not appear to be the case. An alternative approach to inference which might be more robust would follow the marginal regression modelling approach described by Leisenring, Pepe and Longton [17]. One could formulate logistic regression models for the population average sensitivity and 1-specificity as

$$\begin{aligned} \text{logit}\ \{\text{Prob[screen positive} \mid \text{image diseased]}\} &= \gamma_0 + \gamma_1 b \\ \text{logit}\ \{\text{Prob[screen positive} \mid \text{image non-diseased]}\} &= \eta_0 + \eta_1 b \end{aligned}$$

where the logit function is $\text{logit}\{x\} = ln\{x/(1-x)\}$ and b is 0 if the image was read before the intervention and 1 if it was read after the intervention. The changes in the true and false positive rates are now quantified in the odds ratio parameters γ_1 and η_1, respectively, and joint confidence intervals can be calculated. By adding an interaction term between b and I, where I is an indicator of the radiologist being in the control or intervention groups:

$$\begin{aligned}\text{logit}\ \{\text{Prob[screen positive} \mid \text{image diseased]}\} &= \gamma_0 + \gamma_1 b + \gamma_2 bI \\ \text{logit}\ \{\text{Prob[screen positive} \mid \text{image non-diseased]}\} &= \eta_0 + \eta_1 b + \eta_2 bI\end{aligned}$$

a comparison of the changes in the intervention and control groups can be made by testing if the parameters γ_2 or η_2 are 0. Though this logistic regression modelling approach may provide more robust confidence intervals, we felt that the simpler approach described earlier was adequate for power calculations.

The prototype reading study we have described concerns evaluating the effect of an intervention on the change in accuracy parameters. We note, however, that most of our discussion is also relevant to the comparison of accuracies associated with different imaging modalities. Suppose for example, that there are two sets of women (denoted by set 1 and set 2) from which images have been made using two modalities. A natural study design to compare the modalities would entail readers assigned to read one set of films produced with one modality and the other set of films produced with the other modality. Using the notation $1(A)$ to denote set 1 produced with modality A and similarly for the other combination, readers read either $\{1(A)$ and $2(B)\}$ or $\{2(A)$ and $1(B)\}$. Considering that the ordering may also influence accuracy parameters, this yields four groups of readings, $\{1(A),\ 2(B)\}$, $\{2(B),\ 1(A)\}$, $\{2(A),\ 1(B)\}$ and $\{1(B),\ 2(A)\}$. A balanced cross-over design would assign radiologists randomly to these four reading assignments. The difference in the sensitivity and specificity between modality A and B can be calculated by simply pooling all relevant readings for modality A and similarly for modality 2. Inference for the difference follows in the same fashion as that described for the change induced by intervention in the intervention group of our study but that now there are 4 rather than 2 strata of radiologists defined by the image reading set assignments.

Power calculations for reading studies are not straightforward due in part to correlations induced by images and readers. That is, for each image there are multiple readings. Moreover, each reader provides multiple readings and radiologist specific sensitivities and specificities are correlated. We propose valid and simple analyses for dealing with these factors but power calculations required a computer simulation approach. We found the process of developing the computer simulation study to be a useful exercise. It compels one to think through the processes generating study data. It also allows one to experiment with the assumptions and design easily. For example, we considered designs which included a larger number of film sets to be read in the study and found that the study power was decreased slightly due to the extra variation introduced. Computer simulations also allow one to check how test statistics perform under the null hypothesis with sample sizes proposed in the study. Hence one can check if inference based on large sample theory is valid in the setting where it is to be applied. We suggest that simulation studies are a useful approach to power calculations in any setting, though given the complexities in radiology reading studies, the case for the technique in this setting is particularly strong.

ACKNOWLEDGMENTS: We appreciate the support of grants CA63145 and GM54438 awarded by NIH, and grant DAMD17-96-1-6288 awarded by the Department of Defense. We thank Molly Edmonds for her excellent technical help in preparing the

manuscript.

References

1. Miller AB, Chamberlain J, Day NE, Hakama M, Prorok PC. Report on a workshop of the UICC Project on Evaluation of Screening for Cancer. **International Journal of Cancer** 1990; 46: 761-769.

2. Rakowski W, Andersen MR, Stoddard AM, Urban N, Rimer BK, Lane DS, Fox SA, Costanza ME for the NCI Breast Cancer Screening Consortium. A confirmatory analysis of the pros and cons of mammography. **Health Psychology,** in press.

3. Elmore JG, Wells CK, Lee CH, Howard DH and Feinstein AR. Variability in radiologist's interpretation of mammograms. **New England Journal of Medicine** 1994; 3331: 1493-1499.

4. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists: Findings from a national sample. **Archives of Internal Medicine** 1996; 156: 209-213.

5. Begg CB. Advances in statistical methodology for diagnostic medicine in the 1980's. **Statistics in Medicine** 1991; 10: 1887-1895.

6. Begg CB and McNeil BJ. Assessment of radiologic tests: control of bias and other design considerations. **Radiology** 1988; 167: 565-569.

7. Gatsonis C and McNeil BJ. Collaborative evaluations of diagnostic tests: experience of the Radiology Diagnostic Oncology Group. **Radiology** 1990; 175: 571-575.

8. Kopans DB, D'Orsi CJ, Adler DD, Bassett LW, Brenner RJ, Dodd GD, Feig SA, Lopiano MA, McLelland RM, Moskowitz M and Sickles EA. **American College of Radiology Breast Imaging Reporting and Data System, 1993.**

9. Advances in Statistical Methods for Diagnostic Radiology: A Symposium. **Academic Radiology** 1995; 2:S1.

10. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals: rating method data. **Journal of Mathematical Psychology** 1969; 6: 487-496.

11. Swets JA and Pickett RM. **Evaluation of Diagnostic Systems: Methods from Signal Detection Theory.** 1982; Academic Press, New York.

12. Metz CE. ROC methodology in radiologic imaging. **Investigative Radiology** 1986; 21: 720-733.

13. Kopans DB. The accuracy of mammographic interpretations. **New England Journal of Medicine** 1994; 331: 1521-1522.

14. Hanley JA and McNeil BJ. The meaning and use of the area under operating characteristics curve. **Radiology** 1982; 143: 29-36.

15. Johnson RA and Wichern DW. **Applied Multivariate Statistical Analysis.** 1988; Prentice Hall, New Jersey.

16. EGRET Siz module. Statistics and Epidemiology Research Corporation, Seattle, Washington, USA.

17. Leisenring W, Pepe MS and Longton GL. A marginal regression modelling framework for evaluating medical diagnostic tests. **Statistics in Medicine**, (in press).

APPENDIX

1. Variance Estimators for Change in Overall Sensitivity and Specificity.

The change in the overall sensitivity defined in Section 4 can be written formally mathematically as

$$\hat{\Delta}_T(\text{sensitivity}) = \frac{1}{R_T}\left\{ \sum_{r:(\text{order}=1,2)} (\hat{S}_{r,\text{post}} - \hat{S}_{r,\text{pre}}) + \sum_{r:(\text{order}=2,1)} (\hat{S}_{r,\text{post}} - \hat{S}_{r,\text{pre}}) \right\}$$

where $\hat{S}_{r,\text{pre}}$ is the observed sensitivity for radiologist r with his pre-intervention film set and $\hat{S}_{r,\text{post}}$ is the corresponding quantity post-intervention. Observe that the order of film sets essentially defines two strata in this setting and the notation (order $= 1,2$) (or (order $= 2,1$)) used to denote the stratum in the summation indicates that it includes only radiologists assigned sets in the order set 1 first and set 2 second (or set 2 first and set 1 second). The variance of $\hat{\Delta}_T$(sensitivity) can be estimated using the variance of a stratified sample mean $\hat{V} = (\hat{V}_{(1,2)} + \hat{V}_{(2,1)})/R_T$, where $\hat{V}_{(1,2)}$ is the sample variance of the quantities $(\hat{S}_{r,\text{pre}} - \hat{S}_{r,\text{post}})$ in the stratum (order=1,2), and $\hat{V}_2$ is the analogous quantity in the other stratum. The ratio $\hat{\Delta}_T(\text{sensitivity})/\sqrt{\hat{V}}$ can be compared with a standard normal distribution to test for a change in the sensitivity which is statistically significantly different from 0.

2. Chi-squared Test Statistics for Bivariate Analyses.

To simultaneously test the null hypotheses that both the sensitivity and specificity are unchanged in the intervention group, $H_0: \Delta_T(\text{sensitivity}) = 0 = \Delta_T(\text{Specificity})$, the following test statistic can be used

$$\left[\hat{\Delta}_T(\text{sensitivity})\ \hat{\Delta}_T(\text{specificity})\right] \hat{\sum}_T^{-1} \begin{bmatrix} \hat{\Delta}_T(\text{sensitivity}) \\ \hat{\Delta}_T(\text{specificity}) \end{bmatrix}$$

where the square bracket notation is used to denote vectors and $\hat{\sum}_T^{-1}$ is the inverse of a square matrix $\hat{\sum}_T$. This matrix $\hat{\sum}_T$ is a variance-covariance matrix for the two-dimensional statistic $[\hat{\Delta}_T$(sensitivity) $\hat{\Delta}_T$(specificity)], and is the analogue of the variance $\hat{V}$ defined above in relation to the one-dimensional quantity $\hat{\Delta}_T$(sensitivity). Formally we write

$$\hat{\sum}_T = \left\{\hat{\sum}_T^{(1,2)} + \hat{\sum}_T^{(2,1)}\right\} \Big/ (R_T - 1)$$

where $\hat{\sum}_T^{(1,2)}$ is the sample variance-covariance matrix for the quantities $\left\{\hat{S}_{r,\text{post}} - \hat{S}_{r,\text{pre}}\ \hat{F}_{r,\text{post}} - \hat{F}_{r,\text{pre}}\right\}$ in the stratum (order = 1,2), and $\hat{\sum}_T^{2,1}$ is the analogous quantity calculated for the other stratum. The test statistic is compared with a standard chi-squared distribution with 2 degrees of freedom in order to test the null hypothesis concerning changes in sensitivities and specificities.

Consider now the component of the data analysis concerning the comparison of changes between intervention and control groups. Using a subscript C to denote the control group in analogy with our use of the subscript T to denote the intervention group, we define the statistics $\hat{\Delta}_C$(sensitivity), $\hat{\Delta}_C$(specificity) and $\hat{\sum}_C$. The estimated differences between the groups in changes of sensitivities and specificities can be written as $\hat{\Delta}_T(\text{sensitivity}) - \hat{\Delta}_C(\text{sensitivity})$ and $\hat{\Delta}_T(\text{specificity}) - \hat{\Delta}_C(\text{specificity})$, respectively. The hypothesis that the changes are the same for intervention and control groups can be tested by comparing the statistic

$$\left[\hat{\Delta}_T(\text{sens}) - \hat{\Delta}_C(\text{sens})\ \hat{\Delta}_T(\text{spec}) - \hat{\Delta}_C(\text{spec})\right]\ [\textstyle\sum_T + \sum_C]^{-1} \begin{bmatrix} \hat{\Delta}_T(\text{sens}) - \hat{\Delta}_C(\text{sens}) \\ \hat{\Delta}_T(\text{spec}) - \hat{\Delta}_C(\text{spec}) \end{bmatrix}$$

with the quantiles of a chi-squared distribution with 2 degrees of freedom, where we use the abbreviations ‘sens’ and ‘spec’ to denote ‘sensitivity’ and ‘specificity’ in the above expressions.

3. Mixed Models for Reading Accuracies.

Section 5 outlines a statistical model for sensitivity and specificity parameters which vary with reader and image. Here we present a more formal and precise definition of this model. For radiologist r on diseased film i, we write the chance of correctly identifying it as diseased pre-intervention using a logistic model as

$$P^D_{r,i} = \exp\{\mu^D + \gamma^D_i + \beta^D_r\} / \left(1 + \exp\{\mu^D + \gamma^D_i + \beta^D_r\}\right)$$

where γ^D_i and β^D_r are random variables specific to this film and radiologist, respectively. For the average radiologist $\beta^D_r = 0$, and for the average film $\gamma^D_i = 0$. Thus for the average radiologist on the average film the sensitivity is $S^D = \exp\{\mu^D\}/(1 + \exp\{\mu^D\})$. The films vary in difficulty in the sense that the average radiologist has a lower sensitivity on some films and a higher sensitivity on others. Mathematically this translates into allowing γ^D_i to vary. We choose it as a random variable so that the average radiologist's sensitivity to different films varies uniformly in an interval $(S^D - a^D, S^D + a^D)$. Technically this is achieved by letting $\gamma^D_i = \ln\{U^D_i/(1 - U^D_i)\} - \mu^D$ where U^D_i is a random variable with a uniform distribution in $(S^D - a^D, S^D + a^D)$. The radiologists also vary amongst themselves in their sensitivities to the same film and this inter-rater variation translates into allowing β^D_r to vary. We simulated data so that on the average diseased film (i.e., $\gamma^D_i = 0$) the sensitivities of radiologists varied uniformly in $(S^D - b^D, S^D + b^D)$. Again, technically we let $\beta^D_r = \ln\{U^D_r/(1 - U^D_r)\} - \mu^D$ where U^D_r is a random variable with a uniform distribution on the interval $(S^D - b^D, S^D + b^D)$.

Turning now to specificities, we write the specificity for radiologist r on non-diseased film j pre-intervention as

$$P^{\bar{D}}_{r,j} = \exp\{\mu^{\bar{D}} + \gamma_j^{\bar{D}} + \beta_r^{\bar{D}}\}\Big/(1 + \exp\{\mu^{\bar{D}} + \gamma_j^{\bar{D}} + \beta_r^{\bar{D}}\})$$

where in analogy with the above notation for diseased films, the average radiologist on the average film has specificity $S^{\bar{D}} = \exp\{\mu^{\bar{D}}\}/(1 + \exp\{\mu^{\bar{D}})$ and parameters $a^{\bar{D}}$ and b^{D} indicate variation in the specificity with film and radiologist. As argued in section 5, data should be generated so that the β_r^D and $\beta_r^{\bar{D}}$ are negatively correlated. We incorporated this into the simulation by first generating the sensitivity radiologist-specific random effect parameter, β_r^D, (i.e., his/her sensitivity to the average film) which is based on the random variable U_r^D, and then letting the corresponding random variable for the specificity random effect be defined as

$$U_r^{\bar{D}} = \left\{\left(S^{\bar{D}} - (U_r^D - S^D)\frac{b^{\bar{D}}}{b^D}\right)\right\}.$$

Thus if the radiologist's sensitivity is $x \times b^D$ above the average radiologist's sensitivity to the average film, S^D, his/her specificity will be $x \times b^{\bar{D}}$ below the average specificity to the average film.

Our model postulates that after intervention the quantities $S^{\bar{D}}$ and S^D are changed to new values but that the radiologist and image-specific parameters remain unchanged. Thus, suppose that after intervention the sensitivity of the average radiologist to the average film is $\exp(\mu^D + \alpha^D\}/(1 + \exp\{\mu^D + \alpha^D\})$. Then the chances that radiologist r will correctly classify film i pre- and post-intervention are

$$P^D_{r,i,\text{pre}} = \exp\{\mu^D + \gamma_i^D + \beta_r^D\}/(1 + \exp\{\mu^D + \gamma_i^D + \beta_r^D\})$$

and

$$P^D_{r,i,\text{post}} = \exp\{\mu^D + \alpha^D + \gamma_i^D + \beta_r^D\}/(1 + \exp\{\mu^D + \alpha^D + \gamma_i^D + \beta_r^D\}),$$

respectively. Similarly the postulated change in $S^{\bar{D}}$ specifies a parameter $\alpha^{\bar{D}}$ (analogous to α^{D}) which facilitates calculation of post-intervention specificities. Having chosen values for the various parameters $(\mu^{D}, \alpha^{D}, a^{D}, b^{D})$ and $(\mu^{\bar{D}}, \alpha^{\bar{D}}, a^{\bar{D}}, b^{\bar{D}})$, this completes the first step of the simulation power calculation method, namely specification of accuracy parameter distributions pre-intervention and intervention effects.

Table 1. Study power to detect a 10% increase in sensitivity with no accompanying effect on specificity pre- versus post-intervention within the intervention group and power to detect a 10% increase in sensitivity in the intervention group versus no change in the sensitivity in the control group when specificites in both groups remain unchanged after the intervention. All tests are two sided and are tested at a significance level of .05.

				Power	
Readers Per Group (R_T)	Films Per Set (M)	Pre-intervention Sensitivity	Pre-intervention Specificity	Within Intervention Group	Comparison with Control Group
20	30	.70	.70	.70	.38
20	30	.70	.80	.66	.34
20	30	.80	.70	.79	.45
20	30	.80	.80	.77	.44
20	45	.70	.70	.81	.48
20	45	.70	.80	.82	.53
20	45	.80	.70	.91	.61
20	45	.80	.80	.92	.64
30	30	.70	.70	.81	.48
30	30	.70	.80	.83	.52
30	30	.80	.70	.93	.60
30	30	.80	.80	.91	.61
30	45	.70	.70	.94	.66
30	45	.70	.80	.95	.66
30	45	.80	.70	.99	.80
30	45	.80	.80	.99	.79
40	30	.70	.70	.92	.61
40	30	.70	.80	.94	.60
40	30	.80	.70	.97	.73
40	30	.80	.80	.98	.75
40	45	.70	.70	.98	.79
40	45	.70	.80	.99	.80
40	45	.80	.70	.99	.88
40	45	.80	.80	.99	.89

Table 2. Study Power to detect various configurations of changes in the intervention group using a study design with 30 readers and 45 films per set. The pre-intervention specificity is assumed to be .70 in all cases. The intervention induced change in sensitivity is denoted Δ_T(sens) and in specificity is denoted Δ_T(spec).

Pre-Intervention Sensitivity	Δ_T(sens)	Δ_T(spec)	Power
.60	+0.10	0.00	0.90
.70	+0.10	0.00	0.95
.80	+0.10	0.00	0.98
.60	+0.05	0.00	0.35
.70	+0.05	0.00	0.39
.80	+0.05	0.00	0.50
.60	+0.05	+0.05	0.66
.70	+0.05	+0.05	0.68
.80	+0.05	+0.05	0.71

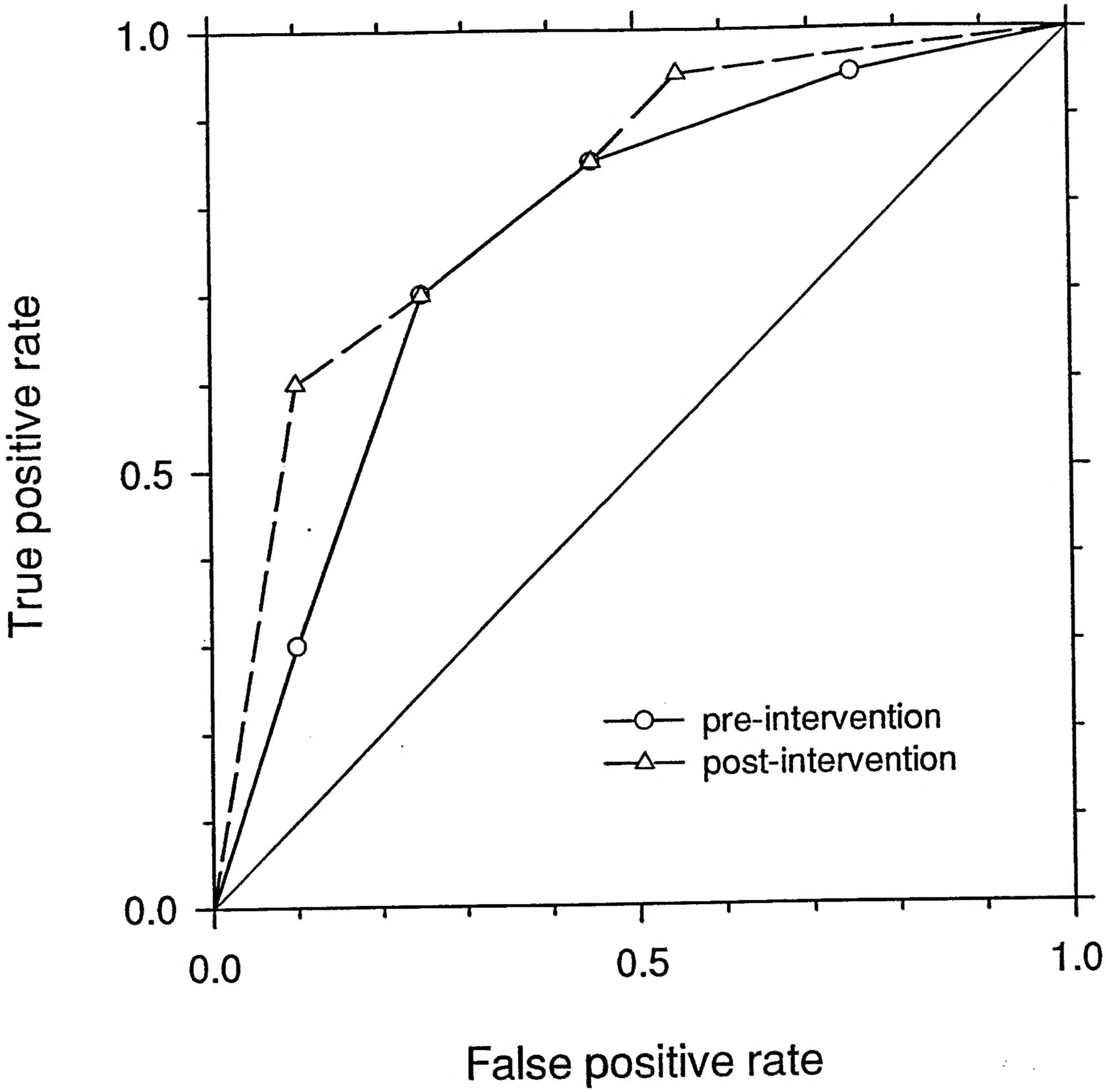

Figure 1: An hypothetical setting where the sensitivity and specificity associated with the clinically relevant criteria are unchanged but the empirical ROC curves indicate a benefit of intervention. The (false positive, true positive) points associated with categories 5, 4, 3, and 2 are (.10, .30), (.25, .70), (.45, .85), and (.75, .95) respectively pre-intervention and (.10, .60), (.25, 70), (.45, .85), and (.55, .95) respectively post-intervention.

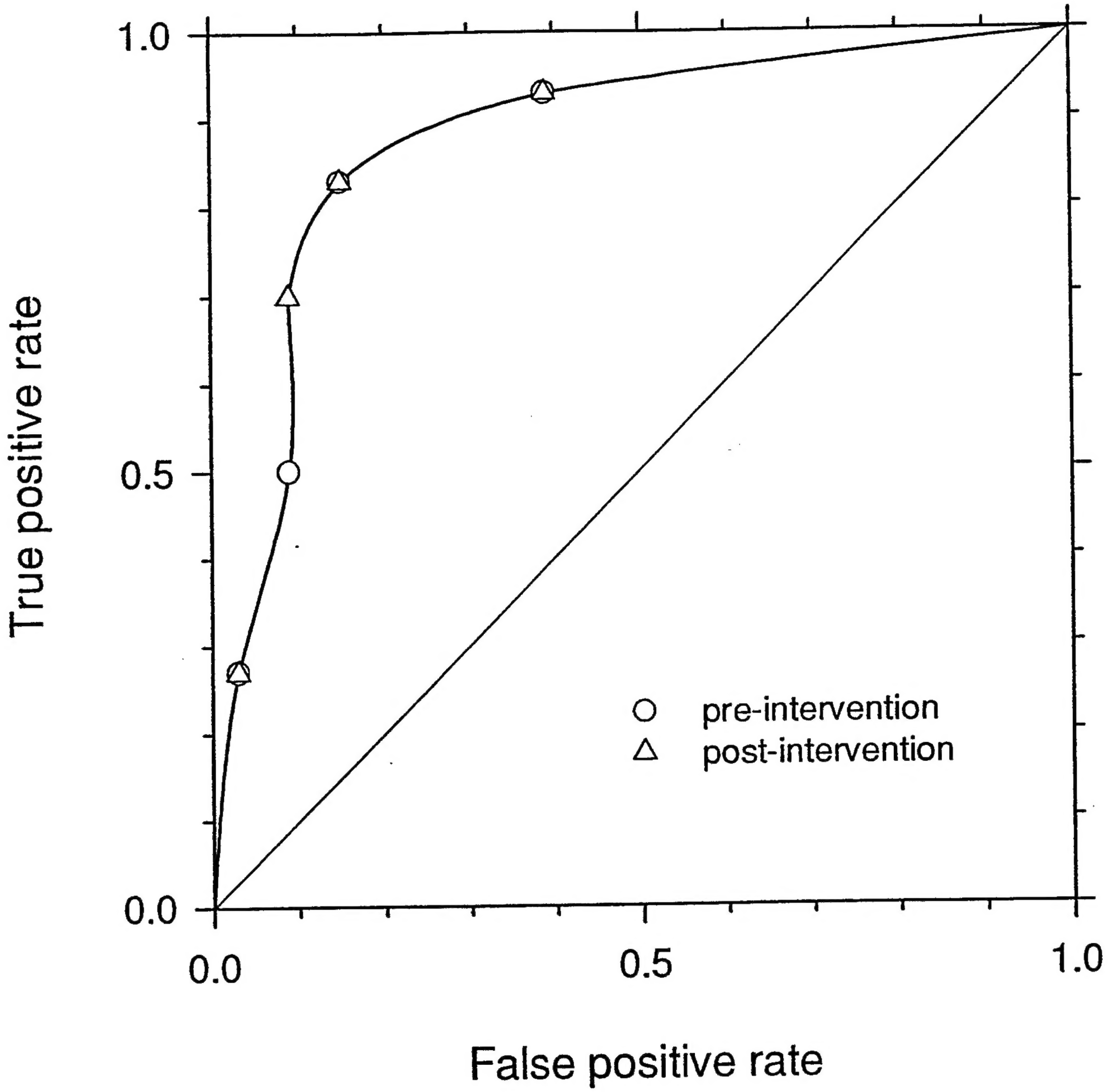

Figure 2: An hypothetical setting where ROC curve is unchanged by the intervention but there is a clinically relevant benefit. The sensitivity associated with the clinically relevant criterion is improved from .50 to .70 while the associated false positive rate remains unchanged at 0.09. The (false positive, true positive) points associated with categories 5, 4, 3, and 2 are (.03, .27), (.09, .50), (.15, .83), and (.39, .93) pre-intervention and (.03, .27), (.09, .70), (.15, .83), and (.39, .93) post-intervention. These points before intervention are labelled with circles and after intervention are labelled with triangles.

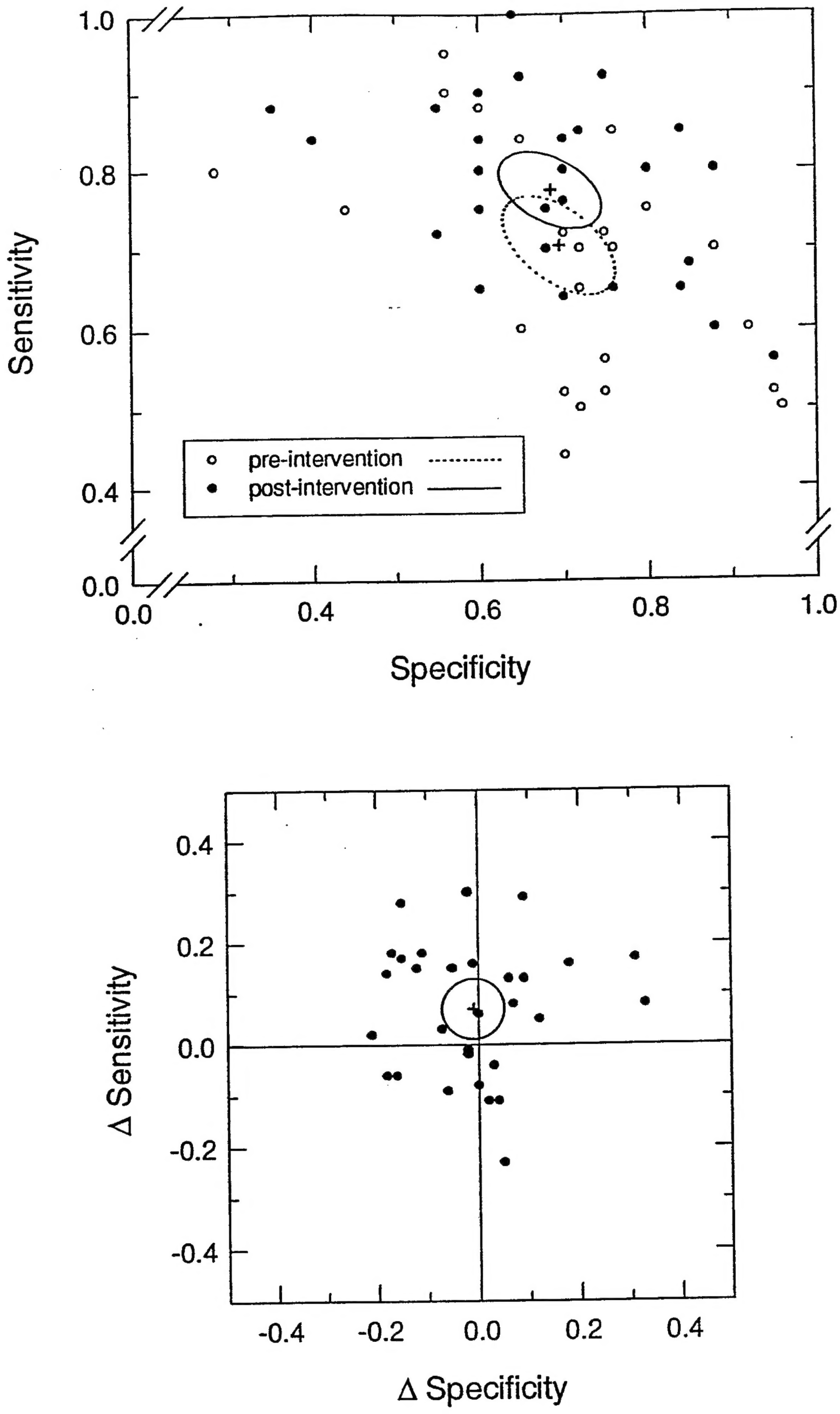

Figure 3: Joint confidence regions for sensitivity and specificity both pre and post intervention along with a joint confidence region for the changes in these parameters. Data used in this illustration were generated using computer simulation methods described in sections 5 and 6. Points correspond to observed data for individual radiologists.